1 Fundamentals

1.1 Optimality conditions

Theorem 1 (Taylor's Theorem)**.** Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable ($f \in \mathcal{C}^1$) and that $d \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$. Then we have that

$$f(x + \alpha d) \approx f(x) + \alpha \nabla f(x)^T d$$

and, if $f \in \mathcal{C}^2$, we also have that

$$f(x + \alpha d) \approx f(x) + \alpha \nabla f(x)^T d + \frac{1}{2}\alpha^2 d^T \nabla^2 f(x) d.$$

Theorem 2 (First Order Necessary Conditions)**.** If x^* is a local minimizer and f is continuously differentiable in an open neighbourhood of x^*, then $\mathbf{\nabla f(x^*) = 0}$.

Theorem 3 (Second Order Necessary Conditions)**.** If x^* is a local minimizer and $\nabla^2 f$ is continuous in an open neighbourhood of x^*, then $\nabla f(x^*) = 0$ and $\mathbf{\nabla^2 f(x^*)}$ **is positive semidefinite**.

Theorem 4 (Second Order Sufficient Conditions)**.** Suppose that $\nabla^2 f$ is continuous in an open neighbourhood of x^* and that $\nabla f(x^*) = 0$ and $\mathbf{\nabla^2 f(x^*)}$ **is positive definite**. Then x^* is a strict local minimizer of f.

Proof. Given that $\nabla^2 f(x^*)$ is positive definite, $d^T \nabla^2 f(x^*) d > 0 \ \forall d \in \mathbb{R}^n$. Hence, we can apply Taylor's Theorem. □

1.2 Descent direction

Definition (Directional derivative)**.** If $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable ($f \in \mathcal{C}^1$) and $d \in \mathbb{R}^n$, then the *directional derivative* of f in the direction d is given by

$$D(f(x); d) \stackrel{\text{def}}{=} \lim_{\epsilon \to 0} \frac{f(x + \epsilon d) - f(x)}{\epsilon} = \nabla f(x)^T d.$$

To verify this formula, we define the function

$$\phi(\alpha) = f(x + \alpha d) = f(y(\alpha)),$$

where $y(\alpha) = x + \alpha d$. Note that

$$\lim_{\epsilon \to 0} \frac{f(x + \epsilon d) - f(x)}{\epsilon} = \lim_{\epsilon \to 0} \frac{\phi(\epsilon) - \phi(0)}{\epsilon} = \phi'(0).$$

By applying the chain rule to $f(y(\alpha))$ we obtain

$$\begin{aligned}\phi'(\alpha) &= \sum_{i=1}^n \frac{\partial f(y(\alpha))}{\partial y_i} \nabla y_i(\alpha) \\ &= \sum_{i=1}^n \frac{\partial f(y(\alpha))}{\partial y_i} d_i \\ &= \nabla f(x + \alpha d)^T d.\end{aligned}$$

Proposition 5. $\nabla f(x)^T d < 0 \Rightarrow d$ is a descent direction for f from x.

1.3 Line Search

Definition. Let f, x and d. Line search is the procedure to find the optimal step length

$$\alpha^* \stackrel{\text{def}}{=} \arg\min_{\alpha > 0} \{\phi(\alpha) = f(x + \alpha d)\}.$$

Proposition 6 (Exact line search)**.** Let $f(x) = \frac{1}{2}x^T Q x - b^T x$ be a convex quadratic function. Then,

$$\alpha^* = -\frac{(Qx - b)^T d}{d^T Q d}.$$

Proof.

$$\begin{aligned}\phi(\alpha) &= \frac{1}{2}(x + \alpha d)^T Q (x + \alpha d) - b^T(x - \alpha d) \\ &= \left(\frac{1}{2} d^T Q d\right)\alpha^2 + \left((x^T Q - b^T) d\right)\alpha + f(x).\end{aligned}$$

Hence,

$$\phi'(\alpha) = 0 \Leftrightarrow \alpha^* = -\frac{(Qx - b)^T d}{d^T Q d}.$$

□

Definition (Wolfe Conditions)**.**

- Sufficient decrease (**WC1**):
$$f(x + \alpha d) \leq f(x) + c_1 \alpha \nabla f(x)^T d$$
- Curvature condition (**WC2**):
$$f(x + \alpha d)^T d \geq c_2 \nabla f(x)^T d$$
$$\phi'(\alpha) \geq c_2 \phi'(0)$$

Definition (Strong Wolfe Conditions)**.**

- Curvature condition (**SWC2**):
$$|f(x + \alpha d)^T d| \leq c_2 |\nabla f(x)^T d|$$

1.4 Global convergence

Definition. An optimization algorithm is said to be globally convergent if $\{x^k\} \underset{k \to \infty}{\longrightarrow} x^*$, i.e. if

$$\lim_{k \to \infty} ||\nabla f(x^k)|| = 0.$$

We will discuss one key property: the angle θ^k between d^k and the steepest descent direction $-\nabla f(x^k)$, defined by:

$$-\nabla f(x^k)^T d^k = ||\nabla f(x^k)|| ||d^k|| \cos \theta^k$$

Theorem 7 (Zoutendijk's Theorem)**.** Consider any iteration of the form $x^k \leftarrow x^k + \alpha^k d^k$, where d^k is a descent direction and α^k satisfies the Wolfe conditions. Suppose that f is bounded below in $\mathbb{R}^n$ in an open set $\mathcal{N}$ containing the level set $\mathcal{L} \stackrel{\text{def}}{=} \{x : f(x) \leq f(x^0)\}$, where x^0 is the starting point of the iteration. Assume also that the gradient ∇f is Lipschitz continuous on $\mathcal{N}$, that is, there exists a constant $L > 0$ such that

$$||\nabla f(x) - \nabla f(\tilde{x})|| \leq ||x - \tilde{x}|| \ \forall x, \tilde{x} \in \mathcal{N}.$$

Then

$$\sum_{k \geq 0} \cos^2 \theta^k \ ||\nabla f(x^k)||^2 < \infty. \tag{1}$$

Inequality (1), which we call the *Zoutendijk condition*, implies that

$$\cos^2\theta^k\ ||\nabla f(x^k)||^2 \to 0.$$

If our method for choosing the search direction d^k ensures that the angle d^k is bounded away from 90^o (*Convergent Angle Condition*), then there is a positive constant δ such that

$$\cos\theta^k \geq \delta > 0\ \forall k.$$

It follows immediately that $\lim_{k\to\infty}||\nabla f(x^k)|| = 0$ and hence the sequence $\{x^k\}$ is convergent.

1.5 Local convergence

Definition. The local convergence of a globally convergent optimization algorithm is the order of convergence of the series $\{x^k\} \underset{k\to\infty}{\longrightarrow} x^*$.

Definition. Let $\{x^k\}$ be a sequence in $\mathbb{R}^n$ that converges to x^*. We say that the convergence is

- **linear** if there is a constant $r \in (0,1)$ such that
$$\frac{||x^{k+1}-x^*||}{||x^k-x^*||} \leq r \text{ for all } k \text{ large enough.}$$
- **superlinear** if
$$\lim_{k\to\infty}\frac{||x^{k+1}-x^*||}{||x^k-x^*||} = 0.$$
- **quadratic** if there is a constant $M > 0$ such that
$$\frac{||x^{k+1}-x^*||}{||x^k-x^*||^2} \leq M \text{ for all } k \text{ large enough.}$$

2 First Derivative Methods

2.1 Gradient Method

$$\boxed{d^k = -\nabla f^k}$$

The Gradient Method is globally convergent as every d^k is a descent direction and $\cos\theta^k = 1\ \forall k$.

Local convergence for quadratic f

Theorem 8. When the gradient method *with exact line searches* is applied to a strongly convex quadratic function $f(x) = \frac{1}{2}x^TQx - b^Tx$, the error norm satisfies

$$f^{k+1} - f^* \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 (f^k - f^*)$$

where $0 < \lambda_1 \leq \cdots \leq \lambda_n$ are the eigenvalues of Q.

Proof. Apply the *Kantorovich inequality* for symmetric positive definite matrices Q:

$$\frac{(x^Tx)^2}{(x^TQx)(x^TQ^{-1}x)} \geq \frac{\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}.$$

□

Local convergence for general nonlinear f

The same results as for quadratic functions hold, substituting Q by $\nabla^2 f^*$.

Proposition 9. If the gradient method is applied with exact line search to a strictly convex quadratic function, $d^k \perp d^{k+1}$. This is not necessarily true if the quadratic function is not strictly convex.

Proof. It comes from the fact that

$$\phi'(\alpha^*) = {\nabla f^{k+1}}^T d^k = 0.$$

□

2.2 Conjugate Gradient Method

$$\boxed{d^k = -\nabla f^k + \beta^k d^{k-1}}$$

The most successful choices for the update coefficient β^k are:

- **Fletcher-Reeves formulae**:
$$\beta^k_{FR} = \frac{||\nabla f^k||}{||\nabla f^{k-1}||^2}$$
- **Polak-Ribière formulae**:
$$\beta^k_{PR} = \frac{{\nabla f^k}^T(\nabla f^k - \nabla f^{k-1})}{||\nabla f^{k-1}||^2}$$

The Fletcher-Reeves formulae has the best theoretical properties whereas the Polak-Ribière is the one showing the best practical behaviour.

Global convergence

The *descent condition* is given by

$${\nabla f^k}^T d^k = -||\nabla f^k||^2 + \beta^k {\nabla f^k}^T d^{k-1} < 0.$$

With **exact line search**, $\alpha^k = \alpha^*$ and ${\nabla f^k}^T d^{k-1} = 0$ so that the previous inequality holds.
With **inexact line search** additional conditions must be imposed to the step length α^k.

- **Fletcher-Reeves**: d^k_{CG-FR} is a descent direction if the step length α^k satisfies SWC with $c_2 < \frac{1}{2}$.
- **Polak-Ribière**: The following modifications are needed to guarantee the descent direction property:
  i. A negative β^k_{PR} must be avoided, taking $\beta^k_{PR+} = \max(0, \beta^k_{PR})$.
  ii. The step length a^k must satisfy WC or SWC and the sufficient decrease condition (SDC):
  $${\nabla f^k}^T d^k \leq -c_3||\nabla f^k||^2,\ 0 < c_3 \leq 1.$$

From a theoretical point of view, the *Convergent Angle Condition* implying $\lim_{k\to\infty}||\nabla f^k|| = 0$ cannot be proved for Conjugate Gradient methods. However, a somehow weaker result can be established:

Theorem 10. Suppose that f satisfies the conditions of Zoutendijk's theorem. If the Conjugate gradient method is applied with either

- β^k_{FR} and α^k satisfying the *SWC* with $c_2 < \frac{1}{2}$, or
- β^k_{PR+} and α^k satisfying the *WC* plus the *SDC*,

then

$$\liminf_{k\to\infty} ||\nabla f^k|| = 0.$$

Local convergence

Restart means to take a gradient step, $d^k = -\nabla f^k$ at every iteration k, namely

- Every n iterations.
- Whenever two consecutive gradients are far from orthogonal:
$$\frac{|\nabla f^{k^T} \nabla f^{k-1}|}{||\nabla f^k||^2} \geq \nu \quad (\nu \approx 0.1).$$

Proposition 11 (n-step Quadratic Convergence)**.** Suppose the Conjugate Gradient Method is applied according to Theorem 10 with restart every n iterations. Then,

$$||x^{k+\mathbf{n}} - x^*|| = \mathcal{O}(||x^k - x^*||^2)$$

2.3 Quasi-Newton Methods

$$\boxed{d^k = -B^{k^{-1}} \nabla f^k}$$

An iteration of a quasi-Newton method sets d^k_{QN} as the minimizer of a quadratic approximation of $f(x)$ around x^k using an **approximation of the true hessian** $B^k \approx \nabla^2 f^k$ **without using second derivatives**.
The most popular expression for the matrix B^k is the **Broyden-Fletcher-Goldfarb-Shanno (BFGS)**.

BFGS Update Formulae

Let $x^{k+1} \leftarrow x^k + \alpha^k d^k_{QN}$.
Let $f^{k+1}_{QN}(d) = f^{k+1} + \nabla f^{k+1^T} d + \frac{1}{2} d^T B^{k+1} d$. How can we make B^{k+1} approximate $\nabla^2 f^{k+1}$ based on the information gathered in the last iteration?

i. First we **impose** f^{k+1}_{QN} **to have the same derivative as** f **at** x^k:
$$\nabla f^{k+1}_{QN}(-\alpha^k d^k_{QN}) = \nabla f^{k+1} - \alpha^k B^{k+1} d^k_{QN} = \nabla f^k$$
$$\nabla f^{k+1} - \nabla f^k = B^{k+1} \alpha^k d^k_{QN}.$$

ii. Defining $s^k = x^{k+1} - x^k$ and $y^k = \nabla f^{k+1} - \nabla f^k$ and $\mathbf{H^k} \overset{\text{def}}{=} \mathbf{B^{k-1}}$, we obtain the **secant equation** *(SE)*
$$B^{k+1} s^k = y^k$$
$$s^k = H^{k+1} y^k.$$

iii. Premultiplying the *(SE)* by y^k we see that in order for $H^{k+1} \succ 0$, s^k and y^k must satisfy the **curvature condition** *(CC)*:
$$y^{k^T} s^k > 0.$$

The secant equation has an infinite number of solutions H^{k+1}.

Definition. Let $\mathbf{H^{k+1}_{BFGS}}$ be the symmetric $n \times n$ matrix satisfying *(SE)* closest to the current matrix H^k with respect to the weighted Frobenius norm.

It can be proved that the unique solution is the *BFGS update formulae*:

$$H^{k+1}_{BFGS} = (I - \rho^k s^k y^{k^T}) H^k_{BFGS} (I - \rho^k y^k s^{k^T}) + \rho^k s^k s^{k^T},$$

with $\rho^k = \frac{1}{y^{k^T} s^k}$.

Proposition 12 (Properties of the BFGS update formulae)**.**

i. H^{k+1}_{BFGS} is symmetric.

ii. H^{k+1}_{BFGS} satisfies the secant equation.

iii. H^{k+1}_{BFGS} is positive definite if α^k satisfies the *WC*.

Finally, the search direction of the BFGS quasi-Newton method is defined as

$$\boxed{d^k_{BFGS} \overset{\text{def}}{=} -H^k_{BFGS} \nabla f^k}$$

Proposition 13. Any step length α^k satisfying the Wolfe conditions guarantees the curvature condition $y^{k^T} s^k > 0$.

Proof. From *(WC2)* we have that
$$\nabla f^{k+1^T} d^k \geq c_2 \nabla f^{k^T} d^k.$$
Since $s^k = x^{k+1} - x^k = \alpha^k d^k$,
$$\nabla f^{k+1^T} \frac{s^k}{\alpha^k} \geq c_2 \nabla f^{k^T} \frac{s^k}{\alpha^k}$$
$$(\nabla f^{k+1} - \nabla f^k)^T s^k \geq (c_2 - 1) \nabla f^{k^T} s^k$$
$$y^k s^k \geq (c_2 - 1) \nabla f^{k^T} s^k$$
$$y^k s^k \geq (c_2 - 1) \alpha^k \nabla f^{k^T} d^k > 0$$
$\square$

Global convergence

It is straightforward to see that the descent condition holds.

Proposition 14. If the matrices $H^k \succ 0$ with an uniformly bounded condition number, i.e. $\kappa(H^k) \leq M$, $M > 0$, $\forall k$, then $\cos\theta^k \geq \frac{1}{M}$.

Local convergence

Theorem 15. Let $\{x^k\}_{k\geq 0}$ be the iterates generated by the BFGS method converging to a minimizer x^* of a function $f \in \mathcal{C}^2$. Suppose that $\nabla^2 f(x^*)$ is Lipschitz continuous and that the sequence $||x^k - x^*|| \underset{k\to\infty}{\longrightarrow} 0$ rapidly enough. Then, $\{x^k\} \underset{k\to\infty}{\longrightarrow} x^*$ **with superlinear order of convergence**.

3 Second Derivative Methods

3.1 Newton's Method

$$\boxed{p^k = -{\nabla^2 f^k}^{-1} \nabla f^k, \ (\alpha^k = 1)}$$

The global convergence of Newton's Method can only be guaranteed if $\nabla^2 f^k \succ 0 \ \forall k$, which ensures that p^k is a descent direction $\forall k$.

Theorem 16. Suppose that $f \in \mathcal{C}^2$ and that the Hessian $\nabla^2 f(x)$ is Lipschitz continuous in a neighbourhood of a solution x^* at which the *SOSC* are satisfied (i.e. $\nabla f^* = 0, \nabla^2 f^* \succ 0$). Consider the sequence $\{x^k\}_{k \geq 0}$ with $x^{k+1} = x^k - {\nabla^2 f^k}^{-1} \nabla f^k$. If the starting point x^0 is sufficiently close to x^*, then:

i. $\{x^k\} \to x^*$.

ii. **The order of convergence of** $\{x^k\}_{k \geq 0}$ **is quadratic**.

iii. The sequence of gradient norms $\{||\nabla f^k||\}_{k \geq 0}$ converges quadratically to zero.

3.2 Modified Newton's Method

$$\boxed{d^k = -(\nabla^2 f^k + E^k)^{-1} \nabla f^k}$$

In order for Newton's Method to become a practical optimization algorithm we must modify the Hessian matrix $B^k \approx \nabla^2 f^k$ ensuring that:

- While x^k is far from x^*, $B^k \succ 0$ so that $d^k = -{B^k}^{-1} \nabla f^k$ is a descent direction $\forall k$, hence acheiving global convergence.
- When x^k is near to x^*, B^k resembles the true Hessian, hence preserving the quadratic order of convergence.

Definition (Condition number)**.** Let A be a symmetric $n \times n$ positive definite matrix with eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. The condition number of A is $\kappa(A) = ||A||_2 ||A^{-1}||_2 = \lambda_n / \lambda_1$.

Global convergence

Theorem 17. Let f be bounded below and twice continuously differentiable in an open set $\mathcal{N}$ containing the level set $\mathcal{L} = \{x | f(x) \leq f(x^0)\}$ where ∇f is Lipschitz continuous. Then, if the Modified Newton Method started at x^0 and the bounded modified condition holds,

$$\kappa(B^k) \leq C, \ C > 0, \ \forall k$$

the algorithm converges to a stationary point, that is

$$\lim_{k \to \infty} \nabla f^k = 0.$$

Proof. As we have seen with the *BFGS* method, if $B^k \succ 0$ and $\kappa(B^k) \leq C, \ C > 0, \ \forall k$, then $\cos \theta^k \geq \frac{1}{C}$. □

Proposition 18. If, in addition to the hypothesis of Theorem 17, $E^* = 0$ at the optimal solution x^*, then $\nabla^2 f^* \succ 0$ and the algorithm converges to a strict local minimizer.

Local convergence

Theorem 19 (Unit step length)**.** Let $f \in \mathcal{C}^2$ in an open set $\mathcal{N}$. Consider the iteration $x^{n+1} \leftarrow x^k + \alpha^k d^k$ with d^k descent direction and α^k satisfying the *WC* with $c_1 < \frac{1}{2}$. If the sequence $\{x^k\}_{k \geq 0}$ converges to a point $x^* \in \mathcal{N}$ such that $\nabla^2 f^* \succ 0$ and

$$\lim_{k \to \infty} \frac{||\nabla f^k + \nabla^2 f^k d^k||}{||d^k||} = 0$$

then $\exists k_0 \geq 0 \mid \alpha^k \in WC \ \forall k \geq k_0$.

Theorem 20. Let $f \in \mathcal{C}^2$ in an open set $\mathcal{N}$. Consider that the Modified Newton Method with $c_1 < \frac{1}{2}$ converges to $x^* \in \mathcal{N}$ and that

i. The *SOSC* are satisfied at x^*.

ii. The Hessian $\nabla^2 f$ is Lipschitz continuous in a neighbourhood of x^*.

iii. $E^k = 0$ for k large enough.

Then the **order of convergence is quadratic**.

Spectral decomposition of $\nabla^2 f^k$

Theorem 21. Let $A \in \mathbb{R}^{n \times n}$, symmetric, then:

i. A has n real eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$.

ii. There exists an orthonormal basis of eigenvectors $Q = [q_1, \ldots, q_n]$.

iii. A diagonalizes, $A = Q \Lambda Q^T$.

Based on the spectral decomposition $\nabla^2 f^k = Q \Lambda Q^T$ we define $B^k \stackrel{\text{def}}{=} Q \tilde{\Lambda} Q^T$ with

$$\tilde{\Lambda} = diag(\max(\delta, \lambda_i)) = \Lambda + \overbrace{diag(\max(0, \delta - \lambda_i))}^{\Delta\Lambda}$$

so that

$$B^k = \overbrace{Q \Lambda Q^T}^{\nabla^2 f^k} + \overbrace{Q (\Delta\Lambda) Q^T}^{E^k}$$

Cholesky factorization of $\nabla^2 f^k$

Theorem 22. Let $A \in \mathbb{R}^{n \times n}$, symmetric and $A \succ 0$. Then there exists a unique upper-triangular matrix R with positive diagonal entries such that $A = R^T R$.

Note. The Cholesky factorization may not exist for a non positive definite Hessian. Moreover, even if it is positive definite, if it is *ill conditioned*, the computation of the factorization can be unstable.

Perhaps the simplest idea is to find a scalar $\tau > 0$ such that $\nabla^2 f^k + \tau I$ is sufficiently positive definite. *Cholesky with Added Multiple of the Identity* follows this approach.

Note. The largest eigenvalue (in absolute value) of $\nabla^2 f^k$ is bounded by the Frobenius norm $||\nabla^2 f^k||_F \stackrel{\text{def}}{=} \sqrt{\sum_i \sum_j (\nabla^2 f^k_{ij})^2}$.