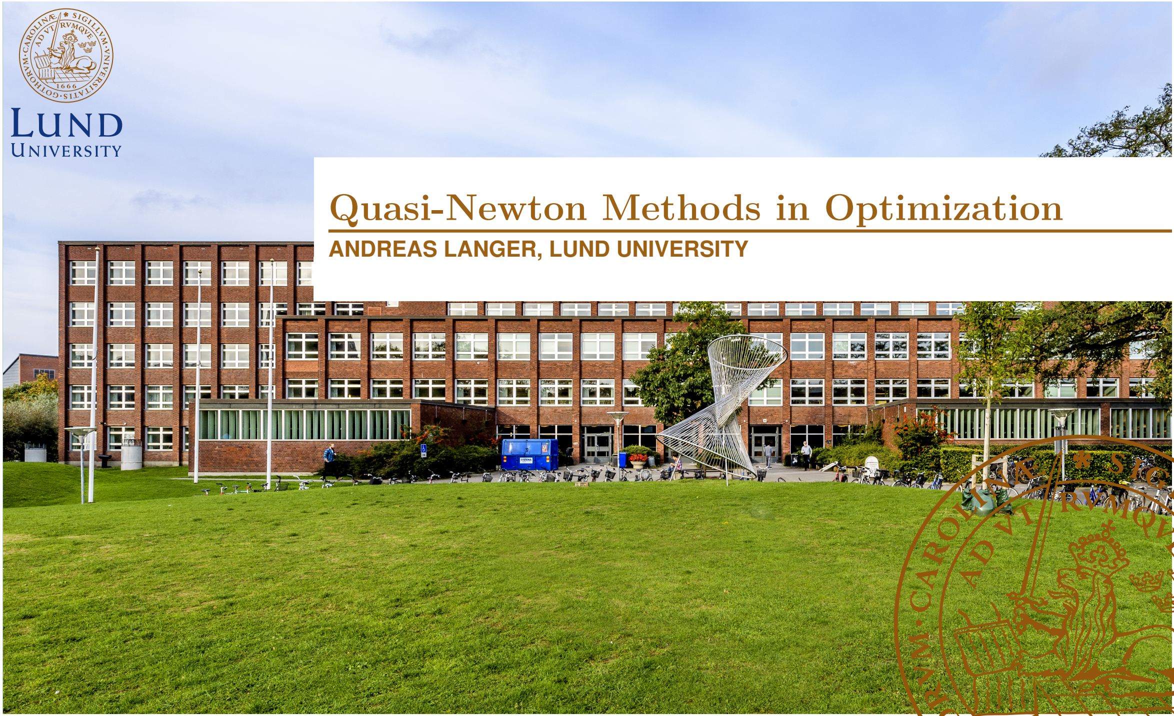
Quasi-Newton Methods in Optimization

ANDREAS LANGER, LUND UNIVERSITY



Problem Description

Let $n \in \mathbb{N}$ and $f : \mathbb{R}^n \to \mathbb{R}$, the *objective function*.

Task: Find $x^* \in \mathbb{R}^n$ such that f is minimal, i.e.,

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

Assumption: f is sufficiently smooth

Characterisation of a Solution

A candidate $x^* \in \mathbb{R}^n$ is called

(i) global minimiser of f in $\mathbb{R}^n$, if

$$f(x^*) \leq f(x) \quad \text{for all } x \in \mathbb{R}^n$$

(ii) strict global minimiser of f in $\mathbb{R}^n$, if

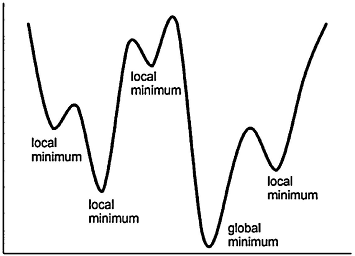
$$f(x^*) < f(x) \quad \text{for all } x \in \mathbb{R}^n \setminus \{x^*\}.$$

(iii) local minimiser of f in $\mathbb{R}^n$, if there exists a neighbourhood of x^* denoted by $N(x^*)$ such that

$$f(x^*) \leq f(x) \quad \text{for all } x \in \mathbb{R}^n \cap N(x^*).$$

(iv) strict local minimiser of f in $\mathbb{R}^n$, if there exists $N(x^*)$ such that

$$f(x^*) < f(x) \quad \text{for all } x \in (\mathbb{R}^n \cap N(x^*)) \setminus \{x^*\}.$$



LUND UNIVERSITY

Existence of a Minimiser

In general a function $f : \mathbb{R}^n \to \mathbb{R}$ does not need to have a minimiser.

Reason: $\mathbb{R}^n$ is **not** compact!

Definition

A subset $X \subseteq \mathbb{R}^n$ is compact, if every sequence in X has a convergent subsequence whose limit belongs to X.

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous. If there exists an $x_0 \in \mathbb{R}^n$ such that the level set

$$\mathcal{L}_f(x_0) := \{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\}$$

is compact, then there exists at least one global minimiser of f in $\mathbb{R}^n$.

Notation

We denote:

- The gradient $\nabla f(x) =: g(x) \in \mathbb{R}^n$ and write it as a row vector.
- The Hessian $\nabla^2 f(x) =: G(x) \in \mathbb{R}^{n\times n}$ (it is a $n \times n$ matrix).

Note: G is a symmetric matrix.

LUND UNIVERSITY

Optimality Conditions

Theorem (Second-order sufficient conditions)

*A candidate $x^* \in \mathbb{R}^n$ is a **local minimiser** of f, if*

- $g(x^*) = 0$ *(first-order necessary optimality condition)*
- *the Hessian $G(x^*)$ is positive definite, i.e., $d^T G(x^*) d > 0$ for all $d \in \mathbb{R}^n \setminus \{0\}$.*

A test for positive definiteness can be made together with Cholesky decomposition. See e.g. `scipy.linalg.chol`

- If f is convex, then any local minimiser of f is also a global minimiser of f.
- If f is strictly convex, then f has at most one local minimiser (which is also a global minimiser).

Numerical Method: Newton

Solve $g(x^*) = 0$ (first-order necessary optimality condition) by iterating:

- Choose an initial value (guess) $x^{(0)} \in \mathbb{R}^n$
- Loop over k until a termination criterion holds:

$$\begin{aligned} s^{(k)} &:= -G(x^{(k)})^{-1} g(x^{(k)}) \\ x^{(k+1)} &:= x^{(k)} + s^{(k)} \end{aligned}$$

We write $g^{(k)} := g(x^{(k)})$ and $G^{(k)} := G(x^{(k)})$.

$s^{(k)} := -G(x^{(k)})^{-1} g(x^{(k)})$ is called the *Newton direction*.

LUND UNIVERSITY

Stopping Criterion

There are basically two termination criteria for Newton's method:

1. **Residual criterion:** The Newton iteration is stopped as soon as the residual $\|g(x^{(k)})\|$ is small enough. In case of convergence we have

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

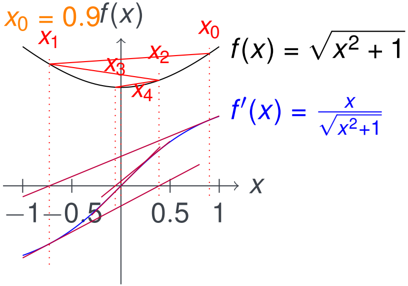
2. **Cauchy criterion:** Terminate the iteration as soon as the Newton-correction $\|x^{(k+1)} - x^{(k)}\| = \|s^{(k)}\|$ is small enough. In case of convergence we have

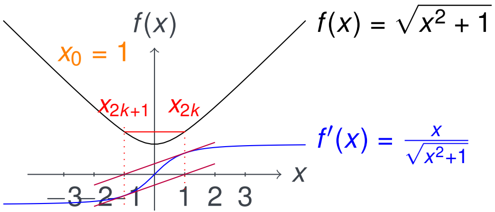
$$\lim_{k\to\infty} \|x^{(k+1)} - x^{(k)}\| = 0.$$

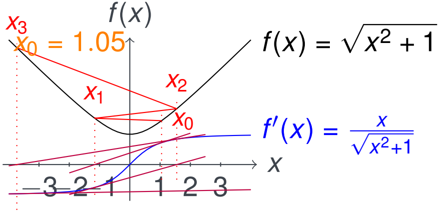
Both criteria should be *used with caution*!

Newton Method: Problem (1)

1st Problem and its remedy:







- Local Convergence: Requires good initial guesses $x^{(0)}$.

Remedy: Globalization $\rightarrow$ *Line search method*

Newton Method: Problem (2)

2nd problem and its remedy:

- Requires the evaluation of the Hessian $G^{(k)}$ in each iteration.
  Remedy: Choose a numerical approximation of $G^{(k)}$ or even better of $(G^{(k)})^{-1}$ $\to$ *Quasi Newton methods*

These two things lead to

$$x^{(k+1)} := x^{(k)} + \alpha^{(k)} s^{(k)}$$

with

$$s^{(k)} := -H^{(k)} g^{(k)}$$

where

- $\alpha^{(k)} > 0$ is a step size;
- $H^{(k)}$ is an approximation of $(G^{(k)})^{-1}$, which can be computed easily.

Line Search

Determine $\alpha^{(k)} > 0$ such that

$$f(x^{(k)} + \alpha^{(k)} s^{(k)}) < f(x^{(k)}).$$

- **Exact line search:** $\alpha^{(k)} \in \arg\min_{\alpha \geq 0} f(x^{(k)} + \alpha s^{(k)})$
- **Inexact line search:** Armijo rule; Powell-Wolfe rule; Goldstein rule; . . . (see e.g. [1],[2] in the course literature)

Define: $\varphi(\alpha) := f(x^{(k)} + \alpha s^{(k)})$

---

[1]: Fletscher, Practical Optimization, 2nd Edition 2013
[2]: Nocedal, Wright: Numerical Optimization, 2006

When does Line Search make sense?

Trivial answer: Whenever there exists an $\alpha \in \mathbb{R}$ such that

$$f(x^{(k)} + \alpha s^{(k)}) < f(x^{(k)}).$$

How can one check that such an α may exist?

Definition (Descent direction)

A direction $s \in \mathbb{R}^n \setminus \{0\}$ is called descent direction of the function $f : \mathbb{R}^n \to \mathbb{R}$ in $x \in \mathbb{R}^n$, if there exists an $\bar{\alpha} > 0$ with

$$f(x + \alpha s) < f(x) \qquad \forall \alpha \in (0, \bar{\alpha}).$$

Descent direction - Criterion

Lemma

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable in $x \in \mathbb{R}^n$. If $s \in \mathbb{R}^n$ fulfils the inequality

$$\nabla f(x)^T s < 0,$$

then s is a descent direction of f in x.

Example (Newton Method): $s^{(k)} = -(G^{(k)})^{-1} g^{(k)} \Rightarrow {g^{(k)}}^T s^{(k)} < 0$ if $G^{(k)}$ positive definite.

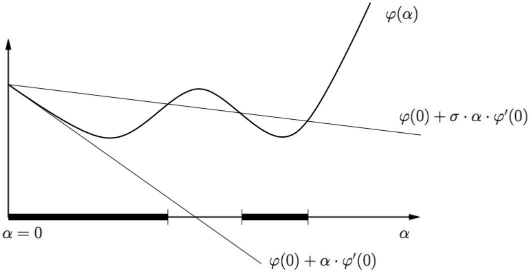
Note: If $s^{(k)}$ is a descent direction of f in $x^{(k)}$, then $\varphi'(0) < 0$.

Armijo Rule

A step size α is acceptable if the following condition

(i) $\varphi(\alpha) \leq \varphi(0) + \sigma\alpha\varphi'(0)$

holds for a given $\sigma \in (0, 1)$.



"This simple and popular strategy for terminating a line search is well suited for Newton methods but is less appropriate for quasi-Newton methods."[2] about Armijo with backtracking

[2]: Nocedal, Wright: Numerical Optimization, 2006

Fundamentals of Optimization

Newton's Method

Line Search

Quasi-Newton Methods

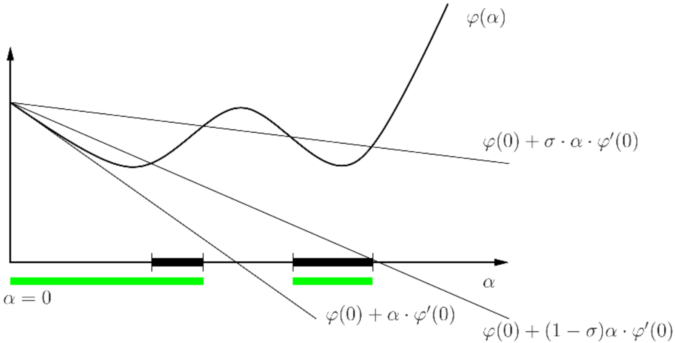
Goldstein Rule

A step size α is acceptable if the following two conditions

(i) $\varphi(\alpha) \leq \varphi(0) + \sigma\alpha\varphi'(0)$ (Armijo rule)

(ii) $\varphi(\alpha) \geq \varphi(0) + (1 - \sigma)\alpha\varphi'(0)$

hold for a given $\sigma \in (0, \frac{1}{2})$.



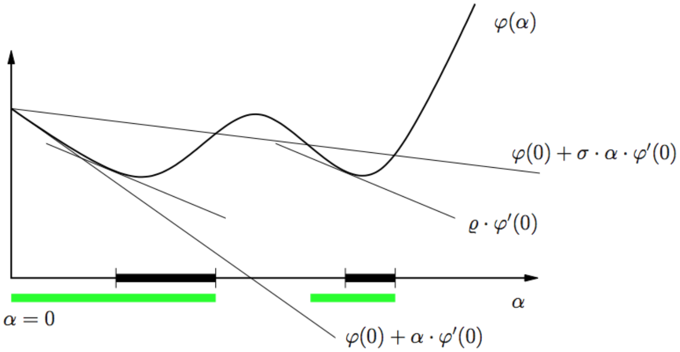
Powell-Wolfe Rule

A step size α is acceptable if the following two conditions

(i) $\varphi(\alpha) \leq \varphi(0) + \sigma\alpha\varphi'(0)$ (Armijo rule)

(ii) $\varphi'(\alpha) \geq \rho\varphi'(0)$

hold for given parameters $\sigma \in (0, \frac{1}{2})$ and $\rho \in (\sigma, 1)$.



Powell-Wolfe Algorithm

Algorithm 1 Powell-Wolfe step size rule

Initialise: Parameters $\sigma \in (0, \frac{1}{2})$, $\rho \in (\sigma, 1)$, $\alpha^- > 0$ (e.g. $\sigma = 10^{-2}$, $\rho = 0.9$, $\alpha^- := 2$)

while α^- does not fulfil (Armijo) condition (i) **do**
  $\alpha^- := \alpha^-/2$
end while
Set $\alpha^+ := \alpha^-$
while α^+ fulfils (Armijo) condition (i) **do**
  $\alpha^+ := 2\alpha^+$
end while
while α^- does not fulfil condition (ii) **do**
  $\alpha_0 := \frac{\alpha^+ + \alpha^-}{2}$
  if α_0 satisfies (Armijo) condition (i) **then**
    $\alpha^- := \alpha_0$
  else
    $\alpha^+ := \alpha_0$
  end if
end while
return $\alpha := \alpha^-$

Quasi-Newton Methods

A typical step looks like:

- Compute $s^{(k)} := -H^{(k)}g^{(k)}$
- Perform line search to compute $\alpha^{(k)}$.
- Compute $x^{(k+1)} := x^{(k)} + \alpha^{(k)}s^{(k)}$
- Update by *some* method $H^{(k)} \rightarrow H^{(k+1)}$

Motivation: Secant Method

In $\mathbb{R}^1$:

$$x^{(k+1)} := x^{(k)} - H^{(k)} g^{(k)}$$

with

$$G^{(k)} \approx \frac{g^{(k)} - g^{(k-1)}}{x^k - x^{k-1}} =: Q^{(k)} = {H^{(k)}}^{-1}$$

In $\mathbb{R}^1$ the approximation $Q^{(k)}$ is **uniquely** determined by

$$Q^{(k)}(x^k - x^{k-1}) = g^{(k)} - g^{(k-1)}. \tag{1}$$

In $\mathbb{R}^n$, $n \geq 2$, is the secant (or Quasi-Newton) condition (1) **NOT uniquely** solvable.
n equations for n^2 unknowns $Q_{ij}^{(k)}$

$\rightarrow$ extra conditions needed.

Broyden condition

Given $Q^{(k)}$. Find $Q^{(k+1)}$ by solving

$$\min \|Q^{(k+1)} - Q^{(k)}\|_{\mathrm{F}}$$

subject to

$$Q^{(k+1)} \underbrace{(x^{k+1} - x^{k})}_{=:\delta^{(k)}} = \underbrace{g^{(k+1)} - g^{(k)}}_{=:\gamma^{(k)}}$$

This gives

$$Q^{(k+1)} = Q^{(k)} + \frac{\gamma^{(k)} - Q^{(k)}\delta^{(k)}}{\delta^{(k)}{}^{\mathrm{T}}\delta^{(k)}}\delta^{(k)}{}^{\mathrm{T}}$$

(see also "good Broyden's method" in `https://en.wikipedia.org/wiki/Broyden%27s_method`)

Sherman – Morrison (Woodbury) formula

Broyden update is a rank-1 update of the form

$$A_1 = A_0 + vw^{\mathrm{T}}$$

where $A_1, A_0 \in \mathbb{R}^{n\times n}$ and $v, w \in \mathbb{R}^n$.

Sherman – Morrison formula gives for the inverse

$$A_1^{-1} = A_0^{-1} - \frac{A_0^{-1} vw^{\mathrm{T}} A_0^{-1}}{1 + w^{\mathrm{T}} A_0^{-1} v}$$

if $1 + w^{\mathrm{T}} A_0^{-1} v \neq 0$.

Simple Rank-1 update

Consider the Sherman – Morrison formula above and replace A_0^{-1} by $H^{(k)}$ and A_1^{-1} by $H^{(k+1)}$.

You then obtain:

$$H^{(k+1)} = H^{(k)} + \frac{(\delta^{(k)} - H^{(k)}\gamma^{(k)})}{{\delta^{(k)}}^{\mathrm{T}} H^{(k)}\gamma^{(k)}} {\delta^{(k)}}^{\mathrm{T}} H^{(k)}$$

Broyden Condition for Inverse Hessian

Alternatively, we might approximate the inverse Jacobian directly by

Find $H^{(k+1)} := {Q^{(k+1)}}^{-1}$ by solving

$$\min \|H^{(k+1)} - H^{(k)}\|_{\mathrm{F}}$$

subject to

$$Q^{(k+1)} \underbrace{(x^{k+1} - x^{k})}_{=:\delta^{(k)}} = \underbrace{g^{(k+1)} - g^{(k)}}_{=:\gamma^{(k)}}$$

This gives

$$H^{(k+1)} = H^{(k)} + \frac{\delta^{(k)} - H^{(k)}\gamma^{(k)}}{{\gamma^{(k)}}^{\mathrm{T}}\gamma^{(k)}}{\gamma^{(k)}}^{\mathrm{T}}$$

(see also "bad Broyden's method" in `https://en.wikipedia.org/wiki/Broyden%27s_method`)

Symmetric Rank 1

Start with a symmetric and invertible matrix $Q^{(k)}$.
Task: Find a symmetric $Q^{(k+1)}$ via a rank-1 update, i.e., of the form

$$Q^{(k+1)} = Q^{(k)} + vw^{\mathrm{T}},$$

fulfilling the Quasi-Newton condition.

This gives

$$Q^{(k+1)} = Q^{(k)} + \frac{(\gamma^{(k)} - Q^{(k)}\delta^{(k)})(\gamma^{(k)} - Q^{(k)}\delta^{(k)})^{\mathrm{T}}}{(\gamma^{(k)} - Q^{(k)}\delta^{(k)})^{\mathrm{T}}\delta^{(k)}}$$

Only rank 1 update which gives a symmetric $Q^{(k+1)}$.
Drawbacks:

- $Q^{(k+1)}$ not necessarily positive definite (even if $Q^{(k)}$ is).
- $(\gamma^{(k)} - Q^{(k)}\delta^{(k)})^{\mathrm{T}}\delta^{(k)} \neq 0$ is not guaranteed.

Symmetric Rank 1 (Inverse)

Consider the Sherman – Morrison formula above and replace A_0^{-1} by $H^{(k)}$ and A_1^{-1} by $H^{(k+1)}$.

You then obtain:

$$H^{(k+1)} = H^{(k)} + auu^{\mathrm{T}}$$

with

$$u := \delta^{(k)} - H^{(k)}\gamma^{(k)}, \quad a := \frac{1}{u^{\mathrm{T}}\gamma^{(k)}}$$

Rank-2 Update – DFP Method

Davidson-Fletcher-Powell (DFP) update

$$Q^{(k+1)} := Q^{(k)} + \left(1 + \frac{\delta^{(k)}{}^{\mathrm{T}} Q^{(k)} \delta^{(k)}}{\gamma^{(k)}{}^{\mathrm{T}} \delta^{(k)}}\right) \frac{\gamma^{(k)} \gamma^{(k)}{}^{\mathrm{T}}}{\gamma^{(k)}{}^{\mathrm{T}} \delta^{(k)}} - \frac{\gamma^{(k)} \delta^{(k)}{}^{\mathrm{T}} Q^{(k)} + Q^{(k)} \delta^{(k)} \gamma^{(k)}{}^{\mathrm{T}}}{\gamma^{(k)}{}^{\mathrm{T}} \delta^{(k)}}$$

$$H^{(k+1)} := H^{(k)} + \frac{\delta^{(k)} \delta^{(k)}{}^{\mathrm{T}}}{\delta^{(k)}{}^{\mathrm{T}} \gamma^{(k)}} - \frac{H^{(k)} \gamma^{(k)} \gamma^{(k)}{}^{\mathrm{T}} H^{(k)}}{\gamma^{(k)}{}^{\mathrm{T}} H^{(k)} \gamma^{(k)}}$$

Rank-2 update - BFGS method

Broyden-Fletcher-Goldfarb-Shanno (BFGS) update

$$Q^{(k+1)} := Q^{(k)} + \frac{\gamma^{(k)} {\gamma^{(k)}}^{\mathrm{T}}}{{\gamma^{(k)}}^{\mathrm{T}} \delta^{(k)}} - \frac{Q^{(k)} \delta^{(k)} {\delta^{(k)}}^{\mathrm{T}} Q^{(k)}}{{\delta^{(k)}}^{\mathrm{T}} Q^{(k)} \delta^{(k)}}$$

$$H^{(k+1)} := H^{(k)} + \left(1 + \frac{{\gamma^{(k)}}^{\mathrm{T}} H^{(k)} \gamma^{(k)}}{{\delta^{(k)}}^{\mathrm{T}} \gamma^{(k)}}\right) \frac{\delta^{(k)} {\delta^{(k)}}^{\mathrm{T}}}{{\delta^{(k)}}^{\mathrm{T}} \gamma^{(k)}} - \frac{\delta^{(k)} {\gamma^{(k)}}^{\mathrm{T}} H^{(k)} + H^{(k)} \gamma^{(k)} {\delta^{(k)}}^{\mathrm{T}}}{{\delta^{(k)}}^{\mathrm{T}} \gamma^{(k)}}$$

see, e.g., Fletcher, R: Practical Methods of Optimization, 2nd Ed, p.55 (reference [1] in the course literature)

BFGS - Curvature condition

Note on BFGS: If ${\gamma^{(k)}}^{\mathrm{T}}\delta^{(k)} > 0$ and $Q^{(k)}$ is positive definite $\Rightarrow$ $Q^{(k+1)}$ is positive definite.

- Armijo rule does not guarantee ${\gamma^{(k)}}^{\mathrm{T}}\delta^{(k)} > 0$.
- But the Powell-Wolfe condition does!!!

Recall secant condition: $Q^{(k+1)}\delta^{(k)} = \gamma^{(k)}$
That a symmetric positive definite matrix $Q^{(k+1)}$ maps $\delta^{(k)}$ into $\gamma^{(k)}$ requires the condition ${\gamma^{(k)}}^{\mathrm{T}}\delta^{(k)} > 0$. (Or in other words: There can only exist a sym. pos. def. matrix $Q^{(k+1)}$ if ${\gamma^{(k)}}^{\mathrm{T}}\delta^{(k)} > 0$).

Powell-Wolfe condition - revisited

Let $\alpha^{(k)}$ be a step-size which fulfils the Powell-Wolfe conditions.

Recall 2nd condition of Powell-Wolfe:

$$\varphi'(\alpha^{(k)}) \geq \rho\varphi'(0) \qquad \text{means}$$
$$\nabla f(\underbrace{x^{(k)} + \alpha^{(k)} s^{(k)}}_{=:x^{(k+1)}})^T s^{(k)} \geq \rho \nabla f(x^{(k)})^T s^{(k)}, \qquad \rho \in (0,1)$$

Recall: $\delta^{(k)} = x^{(k+1)} - x^{(k)} = \alpha^{(k)} s^{(k)}$, $\gamma^{(k)} = \nabla f(x^{(k+1)}) - \nabla f(x^{(k)})$

$$\Rightarrow \nabla f(x^{(k+1)})^T \delta^{(k)} \geq \rho \nabla f(x^{(k)})^T \delta^{(k)} \Rightarrow (\nabla f(x^{(k+1)}) - \rho \nabla f(x^{(k)}))^T \delta^{(k)} \geq 0$$
$$\Rightarrow (\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}))^T \delta^{(k)} \geq (\rho - 1)\alpha^{(k)} \nabla f(x^{(k)})^T s^{(k)}$$
$$\Rightarrow {\gamma^{(k)}}^T \delta^{(k)} > 0$$

as $Q^{(k)}$ is positive definite implying $\nabla f(x^{(k)})^T s^{(k)} < 0$ (cf. descent direction).

LUND UNIVERSITY