Numerical Optimization

Lecture Notes #3
Convergence; Line Search Methods

Fall 2024

Outline

1. Introduction
   - Recap
   - Fundamentals: Rate of Convergence

2. Line Search Methods
   - Search Direction: Steepest Descent, Newton, or Other?!?
   - Step Length Selection — 1D Minimization
   - Step Length Selection — The Wolfe Conditions

Quick Recap: Last Time

Some fundamental building blocks of unconstrained optimization:

Theorem (Taylor)

For some $t \in (0,1)$, we have

$$f(\bar{\mathbf{x}}+\bar{\mathbf{p}}) = f(\bar{\mathbf{x}}) + \bar{\mathbf{p}}^T \underbrace{\nabla f(\bar{\mathbf{x}})}_{gradient} + \frac{1}{2}\bar{\mathbf{p}}^T \underbrace{\left[\nabla^2 f(\bar{\mathbf{x}}+t\bar{\mathbf{p}})\right]}_{Hessian} \bar{\mathbf{p}}.$$

4 theorems relating $f(\bar{\mathbf{x}})$ and its derivatives to optimal solutions.

[1] $\bar{\mathbf{x}}^*$ optimal $\Rightarrow \nabla f(\bar{\mathbf{x}}^*) = 0$.

[2] $\bar{\mathbf{x}}^*$ optimal $\Rightarrow \nabla f(\bar{\mathbf{x}}^*) = 0$, and $\nabla^2 f(\bar{\mathbf{x}}^*)$ positive semi-definite.

[3] $\nabla f(\bar{\mathbf{x}}^*) = 0$, and $\nabla^2 f(\bar{\mathbf{x}}^*)$ positive definite $\Rightarrow \bar{\mathbf{x}}^*$ optimal.

[4a] f convex, and $\bar{\mathbf{x}}^*$ local optimum $\Rightarrow \bar{\mathbf{x}}^*$ global optimum.

[4b] f convex, and $\nabla f(\bar{\mathbf{x}}^*) = 0 \Rightarrow \bar{\mathbf{x}}^*$ global optimum.

Note: The complete statement of the theorems require sufficient smoothness (existence) of derivatives of f.

Definition (Rate of Convergence, Sequences)

Suppose the sequence $\beta = \{\beta_n\}_{n=1}^{\infty}$ converges to zero, and $\underline{\bar{\mathbf{x}}} = \{\bar{\mathbf{x}}_n\}_{n=1}^{\infty}$ converges to a point $\bar{\mathbf{x}}^*$.

If $\exists K > 0$: $\|\bar{\mathbf{x}}_n - \bar{\mathbf{x}}^*\| < K\beta_n$, for $n > N$ (*i.e.* for n large enough), then we say that $\{\bar{\mathbf{x}}_n\}_{n=1}^{\infty}$ converges to $\bar{\mathbf{x}}^*$ with a **Rate of Convergence** $\mathcal{O}(\beta_n)$ ("Big Oh of β_n").

We write

$$\bar{\mathbf{x}}_n = \bar{\mathbf{x}}^* + \mathcal{O}(\beta_n).$$

Note: The sequence $\beta = \{\beta_n\}_{n=1}^{\infty}$ is usually chosen to be *e.g.*

$$\beta_n = n^{-p}, \quad \text{for some value of } p.$$

Rates of Convergence: Example

What does the sequence $1 + (0.5)^k$ converge to? What is the convergence rate?

Let $\bar{\mathbf{x}} = \{\bar{\mathbf{x}}_n\}_{n=1}^{\infty}$ be a sequence converging to $\bar{\mathbf{x}}^*$, the convergence rate is said to be

Q-linear (quotient-linear) if $\exists r \in (0, 1)$ and

$$\frac{\|\bar{\mathbf{x}}_{k+1} - \bar{\mathbf{x}}^*\|}{\|\bar{\mathbf{x}}_k - \bar{\mathbf{x}}^*\|} \leq r, \quad \text{for k sufficiently large}$$

Rates of Convergence

Let $\bar{\mathbf{x}} = \{\bar{\mathbf{x}}_n\}_{n=1}^{\infty}$ be a sequence converging to $\bar{\mathbf{x}}^*$, the convergence rate is said to be

Q-linear (quotient-linear) if $\exists r \in (0,1)$ and

$$\frac{\|\bar{\mathbf{x}}_{k+1} - \bar{\mathbf{x}}^*\|}{\|\bar{\mathbf{x}}_k - \bar{\mathbf{x}}^*\|} \leq r, \quad \text{for k sufficiently large}$$

Q-superlinear if

$$\lim_{k\to\infty} \frac{\|\bar{\mathbf{x}}_{k+1} - \bar{\mathbf{x}}^*\|}{\|\bar{\mathbf{x}}_k - \bar{\mathbf{x}}^*\|} = 0.$$

Q-quadratic if $\exists M \in \mathbb{R}^+$ and

$$\frac{\|\bar{\mathbf{x}}_{k+1} - \bar{\mathbf{x}}^*\|}{\|\bar{\mathbf{x}}_k - \bar{\mathbf{x}}^*\|^2} \leq M, \quad \text{for k sufficiently large}$$

Line Search Methods

Consider that we want to solve $\min_{\bar{\mathbf{x}} \in \mathbf{R}^n} f(\bar{\mathbf{x}})$.

We can reduce this n-dimensional problem to a one-dimensional problem that can be solved iteratively via a line search method. Key steps for a line search method: *(i)* pick a **search direction** $\bar{\mathbf{p}}_k$ and, then *(ii)* solve the one-dimensional problem

$$\min_{\alpha_k > 0} f(\bar{\mathbf{x}}_k + \alpha_k \bar{\mathbf{p}}_k).$$

The solution gives us an optimal value for α_k, so the next point is given by

$$\bar{\mathbf{x}}_{k+1} = \bar{\mathbf{x}}_k + \alpha_k \bar{\mathbf{p}}_k,$$

where α_k is known as the **step length**. In order for a line search method to work well, we need good choices of the direction $\bar{\mathbf{p}}_k$ and the step length α_k.

How we choose $\bar{\mathbf{p}}_k$

We can choose $\bar{\mathbf{p}}_k$ for a line search method by using

- the **Steepest Descent Method** or
- the **Newton Method**.

In the next couple of slides, we will derive the Steepest Descent Method and the Newton method by using Taylor expansion.

Note that once we obtain $\bar{\mathbf{p}}_k$ we can solve the one-dimensional problem

$$\min_{\alpha_k > 0} f(\bar{\mathbf{x}}_k + \alpha_k \bar{\mathbf{p}}_k).$$

to obtain the optimal α_k. Given a starting point $\bar{\mathbf{x}}_0$, the subsequent points can be computed iterative:

$$\bar{\mathbf{x}}_{k+1} = \bar{\mathbf{x}}_k + \alpha_k \bar{\mathbf{p}}_k.$$

Steepest Descent Direction

The intuitive choice for $\bar{\mathbf{p}}_k$ is to move in the direction of steepest descent, *i.e.* in the negative gradient direction.

Going back to the Taylor expansion

$$f(\bar{\mathbf{x}} + \alpha\bar{\mathbf{p}}) = f(\bar{\mathbf{x}}) + \alpha\bar{\mathbf{p}}^T\nabla f(\bar{\mathbf{x}}),$$

we immediately see that the direction of most rapid decrease gives

$$\min_{\|\bar{\mathbf{p}}\|=1} \bar{\mathbf{p}}^T\nabla f(\bar{\mathbf{x}}) = \min_{\theta\in[0,2\pi]} \cos\theta\,\|\nabla f(\bar{\mathbf{x}})\| = -\|\nabla f(\bar{\mathbf{x}})\|,$$

which is achieved when $\theta = \pi \Leftrightarrow \bar{\mathbf{p}} = -\nabla f(\bar{\mathbf{x}})/\|\nabla f(\bar{\mathbf{x}})\|$.

Recall: $\bar{\mathbf{v}}^T\bar{\mathbf{w}} = \cos\theta\,\|\bar{\mathbf{v}}\|\cdot\|\bar{\mathbf{w}}\|$, where θ is the angle between the vectors $\bar{\mathbf{v}}$ and $\bar{\mathbf{w}}$.

Steepest Descent Direction

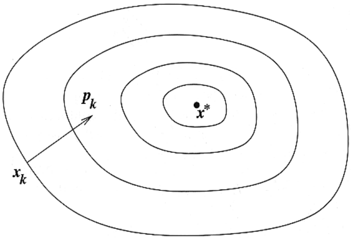


Figure: The steepest descent direction $\bar{\mathbf{p}}_k$ is perpendicular to the contour lines of the objective.

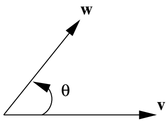


Figure: $\bar{\mathbf{v}}^T\bar{\mathbf{w}} = \cos\theta\, \|\bar{\mathbf{v}}\| \cdot \|\bar{\mathbf{w}}\|$.

Newton Direction

If f is smooth enough and the Hessian is positive definite, we can select $\bar{\mathbf{p}}_k$ to be the "Newton direction." We write down the second order Taylor expansion:

$$f(\bar{\mathbf{x}} + \bar{\mathbf{p}}) \approx f(\bar{\mathbf{x}}) + \bar{\mathbf{p}}^T \nabla f(\bar{\mathbf{x}}) + \frac{1}{2}\bar{\mathbf{p}}^T \left[\nabla^2 f(\bar{\mathbf{x}})\right] \bar{\mathbf{p}}.$$

We seek the minimum of the right-hand-side by computing the derivative width respect to $\bar{\mathbf{p}}$ and set the result to zero

$$\nabla f(\bar{\mathbf{x}}) + \left[\nabla^2 f(\bar{\mathbf{x}})\right] \bar{\mathbf{p}} = 0,$$

which gives the Newton direction

$$\bar{\mathbf{p}}^N = -\left[\nabla^2 f(\bar{\mathbf{x}})\right]^{-1} \nabla f(\bar{\mathbf{x}}).$$

Newton Direction

Recall Taylor expansion: $f(\bar{\mathbf{x}} + \alpha\bar{\mathbf{p}}) = f(\bar{\mathbf{x}}) + \alpha\bar{\mathbf{p}}^T \nabla f(\bar{\mathbf{x}})$,

As long as the Hessian is positive definite, $\bar{\mathbf{p}}^N$ is a descent-direction:

$$\bar{\mathbf{p}}^N \nabla f(\bar{\mathbf{x}}) = -\nabla f(\bar{\mathbf{x}})^T \underbrace{\left[\nabla^2 f(\bar{\mathbf{x}})\right]^{-1}}_{\text{Pos. Def.}} \nabla f(\bar{\mathbf{x}}) < 0$$

Note: Clearly, the Newton direction is more "expensive" than the steepest descent direction — we must compute the Hessian matrix $\nabla^2 f(\bar{\mathbf{x}})$, and invert it (*i.e.* solve an $n \times n$ linear system).

Note: The convergence rate for steepest descent methods is **linear** and for Newton methods it is **quadratic**, hence there is a lot to gain by finding the Newton direction.

Example: NW$^{\text{1st}}$-2.2, p 30.

Problem: Show that the function $f(x) = 8x+12y+x^2-2y^2$ has only one stationary point, and that it is neither a maximum nor a minimum, but a saddle point. Sketch the contours for f.

Solution: The gradient of f is

$$\nabla f = \begin{bmatrix} 8+2x \\ 12-4y \end{bmatrix}$$

which has the stationary point $(x,y) = (-4,3)$. Since the Hessian

$$\nabla^2 f = \begin{bmatrix} 2 & 0 \\ 0 & -4 \end{bmatrix}$$

has both positive and negative eigenvalues, the stationary point must be a saddle point.

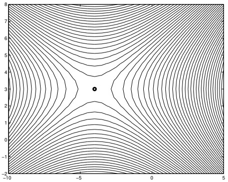


Figure: The contour lines for $f(x)$.

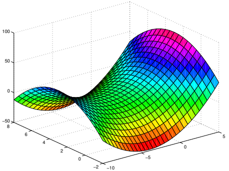


Figure: The function $f(x)$ around the stationary point.

Example: NW$^{\text{1st}}$-2.2, p 30.

If we start an iteration in $(x_0, y_0) = (0, 0)$:

The steepest descent direction is

$$\bar{\mathbf{p}}_0^{\text{SD}} = -\nabla f = -\begin{bmatrix} 8 + 2x \\ 12 - 4y \end{bmatrix} = -\begin{bmatrix} 8 \\ 12 \end{bmatrix}$$

and the Newton direction is

$$\bar{\mathbf{p}}_0^{N} = -[\nabla^2 f]^{-1} \nabla f = -\begin{bmatrix} 2 & 0 \\ 0 & -4 \end{bmatrix}^{-1} \begin{bmatrix} 8 \\ 12 \end{bmatrix} = \begin{bmatrix} -4 \\ 3 \end{bmatrix}$$

Example: NW^{1st}-2.2, p 30.

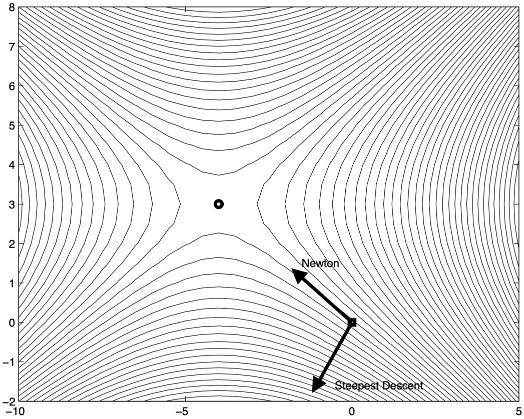


Figure: The **Newton** and **Steepest Descent** directions starting in $(0, 0)$. Note that the Newton method is heading to the saddle point, but the Steepest descent method will, in general, not converge to a non-minimum stationary point.

Modified (Convexified) Example

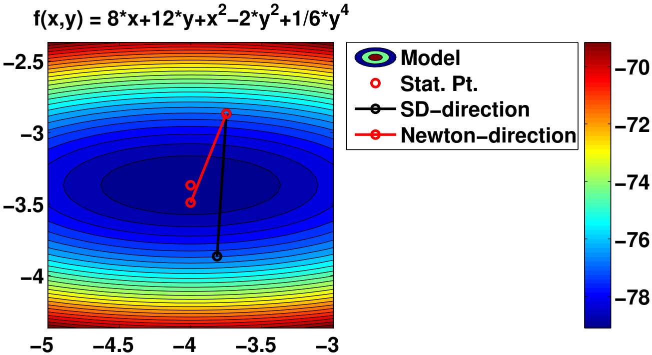


Figure: Convexification of the silly book problem. Same point of interest, $\nabla f = [8 + 2x,\ 12 - 4y + 2/3y^3]^T$, $\nabla^2 f = \begin{bmatrix} 2 & 0 \\ 0 & -4 + 2y^2 \end{bmatrix}$. Now, both the steepest descent and Newton directions are descent directions.

Line Search Methods — Directions

Method	Search Direction	Convergence
Steepest Descent	$p_k = -\nabla f(\bar{\mathbf{x}}_k)/\|\nabla f(\bar{\mathbf{x}}_k)\|$	Linear
Quasi-Newton	$p_k = -H_k^{-1}\,\nabla f(\bar{\mathbf{x}}_k)$	Super-Linear
Newton	$p_k = -[\nabla^2 f(\bar{\mathbf{x}}_k)]^{-1}\,\nabla f(\bar{\mathbf{x}}_k)$	Quadratic

Table: Summary of search directions for different schemes. In **Quasi-Newton** schemes we do not explicitly compute the Hessian $\nabla^2 f(\bar{\mathbf{x}}_k)$ in each iteration, instead we use an approximation $H_k \approx \nabla^2 f(\bar{\mathbf{x}}_k)$ which is updated in some clever way [TO BE EXPLORED IN GREAT DETAIL LATER] (lecture 18→ ...).

We will return to the selection of $\bar{\mathbf{p}}_k$, but let's consider the computation of the step length α_k...

Line Search Methods: Step Length Selection

Given a descent direction $\bar{\mathbf{p}}_k$ we would like to find the global minimizer α_k^* of

$$\min_{\alpha_k > 0} f(\bar{\mathbf{x}}_k + \alpha_k \bar{\mathbf{p}}_k).$$

As this is just one of possible many steps in the iteration, it is not wise to expend too much time in finding α_k. We are faced with a trade-off:

— We want an α_k so that we get a **substantial reduction** in the objective f.

— We want to find α_k **fast**.

In practice we perform an **inexact line search** — settling for an α_k which gives **adequate reduction** in the objective.

What is "adequate reduction?"

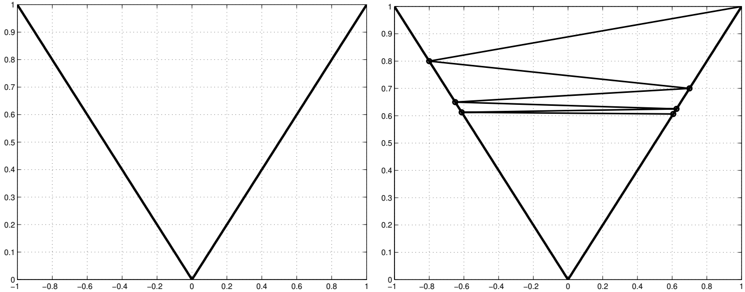


Figure: Consider the objective $f(x) = \sqrt{x^2 + 10^{-8}}$, if we let $x_k = \{1, -0.8, 0.7, -0.65, 0.625, -0.6125, 0.60625, \ldots\}$, then the descent directions are given by $p_k = \{-1, 1, -1, 1, -1, 1, -1, \ldots\}$, so this generates a decreasing sequence $f(x_k + \alpha_k p_k) < f(x_k)$. However, with the current choice of $\alpha_k = \{1.8, -1.5, 1.35, -1.275, 1.2375, -1.21875, \ldots\}$ the convergence rate is less than spectacular. Note that y-axis represents $f(x)$ and x-axis represents x on both graphs.

Clearly, we need a stronger condition than $f(x_k + \alpha_k p_k) < f(x_k)$.

The Wolfe Conditions

There are many ways to enforce reduction in the objective, *e.g.*

Armijo Condition (Wolfe Condition #1)

The **Armijo Condition**

$$f(\bar{\mathbf{x}}_k + \alpha \bar{\mathbf{p}}_k) \leq f(\bar{\mathbf{x}}_k) + c_1 \alpha \bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}), \quad c_1 \in (0, 1),$$

requires the reduction to be proportional to the step length α, as well as the directional derivative $\bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}})$. **In practice** c_1 is usually set to be quite small, *e.g.* $\sim 10^{-4}$.

Armijo Condition requires that the step length cause a sufficient decrease in the objective function value.

The Wolfe Conditions

To rule out unacceptably short steps, we additionally enforce

Curvature Coondition (Wolfe Condition #2)

The **Curvature Condition**

$$\bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}_k + \alpha \bar{\mathbf{p}}_k) \geq c_2 \bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}_k), \quad c_2 \in (c_1, 1).$$

It prevents us from stopping when more progress can be made by moving further (increasing α).

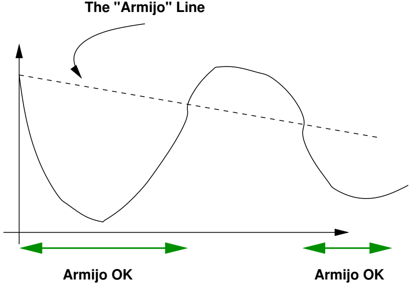
Together these two conditions are known as the **Wolfe conditions**.

The Wolfe Conditions: Part I — The Armijo Condition

The **Armijo Condition**

$$f(\bar{\mathbf{x}}_k + \alpha \bar{\mathbf{p}}_k) \leq f(\bar{\mathbf{x}}_k) + c_1 \alpha \bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}), \quad c_1 \in (0, 1)$$

requires the reduction to be proportional to the step length α, as well as the directional derivative. **In practice** c_1 is usually set to be quite small, *e.g.* $\sim 10^{-4}$.



Here y-axis represents $f(x)$ and x-axis represents α.

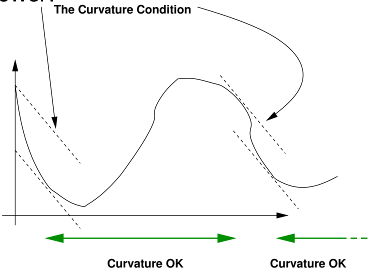
The Wolfe Conditions: Part II — The Curvature Condition

To rule out unacceptable short steps, the **curvature condition**

$$\bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}_k + \alpha \bar{\mathbf{p}}_k) \geq c_2 \bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}_k), \quad c_2 \in (c_1, 1)$$

— it prevents us from stopping when more progress can be made by moving further (increasing α). Typical values: $c_2^{\text{N,QN}} = 0.9$, $c_2^{\text{CG}} = 0.1$.

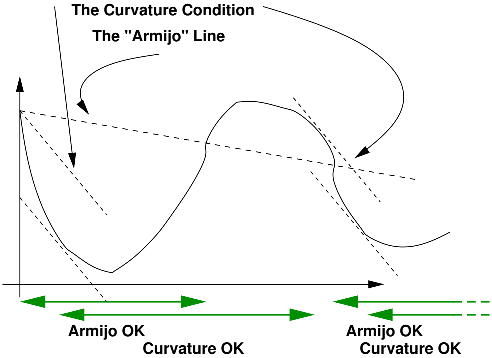
— The curvature condition requires that the directional derivative at the next iterate be shallower.



Here y-axis represents $f(x)$ and x-axis represents α.

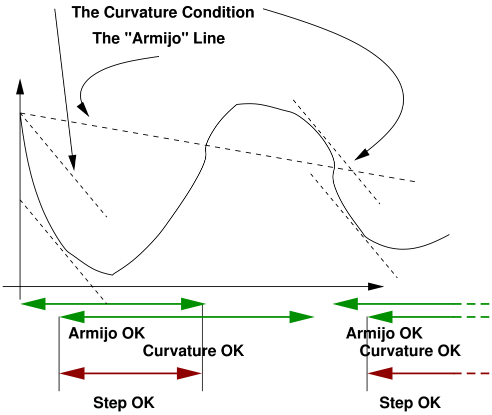
The Wolfe Conditions: Part I+II — Acceptable Step

Together, the Armijo and Curvature conditions constitute the Wolfe Conditions.



The Wolfe Conditions: Part I+II — Acceptable Step

Together, the Armijo and Curvature conditions constitute the Wolfe Conditions.



The Strong Wolfe Conditions

A step length α may satisfy the **Wolfe Conditions**

$$\begin{array}{rcll} f(\bar{\mathbf{x}}_k + \alpha \bar{\mathbf{p}}_k) & \leq & f(\bar{\mathbf{x}}_k) + c_1 \alpha \bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}), & c_1 \in (0,1) \\ \bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}_k + \alpha \bar{\mathbf{p}}_k) & \geq & c_2 \bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}_k), & c_2 \in (c_1, 1) \end{array}$$

even though it is far from a minimizer of $f(\bar{\mathbf{x}}_k + \alpha \bar{\mathbf{p}}_k)$, the **Strong Wolfe Conditions**

$$\begin{array}{rcll} f(\bar{\mathbf{x}}_k + \alpha \bar{\mathbf{p}}_k) & \leq & f(\bar{\mathbf{x}}_k) + c_1 \alpha \bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}), & c_1 \in (0,1) \\ |\bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}_k + \alpha \bar{\mathbf{p}}_k)| & \leq & c_2 |\bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}_k)|, & c_2 \in (c_1, 1) \end{array}$$

further disallows values of

$$\left[\bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}_k + \alpha \bar{\mathbf{p}}_k) \right]$$

which are "too positive," thus excluding point that are far from the stationary points of $\bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}_k + \alpha \bar{\mathbf{p}}_k)$.

Are the Wolfe Conditions too Restrictive?

It can be shown (see NW^{2nd} pp.35–36) that there **exist** step lengths α which satisfy the Wolfe Conditions (and the Strong Wolfe Conditions) **for every** function f which is smooth and bounded below.

Formally —

Theorem (Existence of Acceptable α)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Let $\bar{\mathbf{p}}_k$ be a descent direction at $\bar{\mathbf{x}}_k$, and assume that f is bounded below along the line $\{\bar{\mathbf{x}}_k + \alpha \bar{\mathbf{p}}_k \,:\, \alpha > 0\}$. Then if $0 < c_1 < c_2 < 1$, there exist intervals of step lengths satisfying the Wolfe conditions and the strong Wolfe conditions.

See also "Goldstein Conditions" (NW^{2nd} p.36.)

Algorithm: Backtracking Linesearch

Algorithm: Backtracking Linesearch

[0] `Find a descent direction` $\bar{\mathbf{p}}_k$
[1] `Set` $\overline{\alpha} > 0$, $\rho \in (0, 1)$, $c \in (0, 1)$, `set` $\alpha = \overline{\alpha}$
[2] `While` $f(\bar{\mathbf{x}}_k + \alpha \bar{\mathbf{p}}_k) > f(\bar{\mathbf{x}}_k) + c\alpha \bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}_k)$
[3] $\quad \alpha = \rho\alpha$
[4] `End-While`
[5] `Set` $\alpha_k = \alpha$

If an algorithm selects the step lengths appropriately (*e.g.* backtracking), we do not have to check the second inequality of the Wolfe conditions.

The algorithm above is especially well suited for use with Newton method ($\bar{\mathbf{p}}_k = \bar{\mathbf{p}}_k^N$), where $\overline{\alpha} = 1$. It is less successful for quasi-Newton and CG-based approaches.

The value of the **contraction factor** ρ can be allowed to vary at each iteration of the line search. (To be revisited)

Index