Numerical Optimization

Lecture Notes #3
— Convergence; Line Search Methods —

Peter Blomgren,
⟨blomgren.peter@gmail.com⟩

Department of Mathematics and Statistics
Dynamical Systems Group
Computational Sciences Research Center
San Diego State University
San Diego, CA 92182-7720

http://terminus.sdsu.edu/

Fall 2017
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
   - Homework #1
Some fundamental building blocks of unconstrained optimization:

### Theorem (Taylor)

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

$$f(\bar{x} + \bar{p}) = f(\bar{x}) + \bar{p}^T \nabla f(\bar{x}) + \frac{1}{2} \bar{p}^T \nabla^2 f(\bar{x} + t\bar{p}) \bar{p}.$$  

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

1. $\bar{x}^*$ optimal $\Rightarrow$ $\nabla f(\bar{x}^*) = 0$.
2. $\bar{x}^*$ optimal $\Rightarrow$ $\nabla f(\bar{x}^*) = 0$, and $\nabla^2 f(\bar{x}^*)$ positive semi-definite.
3. $\nabla f(\bar{x}^*) = 0$, and $\nabla^2 f(\bar{x}^*)$ positive definite $\Rightarrow$ $\bar{x}^*$ optimal.
4a. $f$ convex, and $\bar{x}^*$ local optimum $\Rightarrow$ $\bar{x}^*$ global optimum.
4b. $f$ convex, and $\nabla f(\bar{x}^*) = 0$ $\Rightarrow$ $\bar{x}^*$ global optimum.

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

---

Peter Blomgren, {blomgren.peter@gmail.com}
Definition (Rate of Convergence, Sequences)

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

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

We write

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

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

$$\beta_n = \frac{1}{n^p}, \text{ for some value of } p.$$
Rates of Convergence

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

**Q-linear** (quotient-linear) if $\exists r \in (0, 1)$ and $K \in \mathbb{Z}$ such that

$$\frac{\|\bar{x}_{k+1} - \bar{x}^*\|}{\|\bar{x}_k - \bar{x}^*\|} \leq r, \quad \forall k \geq K.$$

**Q-superlinear** if

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

**Q-quadratic** if $\exists r \in \mathbb{R}^+$ and $K \in \mathbb{Z}$ such that

$$\frac{\|\bar{x}_{k+1} - \bar{x}^*\|}{\|\bar{x}_k - \bar{x}^*\|^2} \leq r, \quad \forall k \geq K.$$
Line Search Methods

We now focus on line search methods where we (i) pick a search direction $\bar{p}_k$ and, then (ii) solve the one-dimensional problem

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

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

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

where $\alpha_k$ is known as the step length.

In order for a line search method to be work well, we need good choices of the direction $\bar{p}_k$ and the step length $\alpha_k$. 
Steepest Descent Direction

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

Going back to the Taylor expansion

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

we immediately see that the direction of most rapid decrease gives

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

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

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

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

$$\bar{v}^T\bar{w} = \cos \theta \|\bar{v}\| \cdot \|\bar{w}\|.$$
If $f$ is smooth enough and the Hessian is positive definite, we can select $\bar{p}_k$ to be the “Newton direction.” We write down the second order Taylor expansion:

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

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

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

which gives the Newton direction

$$\bar{p}^N = - \left[ \nabla^2 f(\bar{x}) \right]^{-1} \nabla f(\bar{x}).$$
As long as the Hessian is positive definite, $\bar{p}^N$ is a descent-direction:

$$\bar{p}^N \nabla f(\bar{x}) = -\nabla f(\bar{x})^T \left[ \nabla^2 f(\bar{x}) \right]^{-T} \nabla f(\bar{x}) < 0$$

Note: Clearly, the Newton direction is more “expensive” than the steepest descent direction — we must compute the Hessian matrix $\nabla^2 f(\bar{x})$, 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.
**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.
If we start an iteration in \((x_0, y_0) = (0, 0)\):

The steepest descent direction is

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

and the Newton direction is

\[
\bar{p}^N_0 = -[\nabla^2 f]^{-1} \nabla f = - \begin{bmatrix} 2 & 0 \\ 0 & -4 \end{bmatrix}^{-1} \begin{bmatrix} 8 \\ 12 \end{bmatrix} = \begin{bmatrix} -4 \\ 3 \end{bmatrix}
\]
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.

Peter Blomgren, ⟨blomgren.peter@gmail.com⟩
Modified (Convexified) Example

\[ f(x, y) = 8x + 12y + x^2 - 2y^2 + \frac{1}{6}y^4 \]

**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

<table>
<thead>
<tr>
<th>Method</th>
<th>Search Direction</th>
<th>Convergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steepest Descent</td>
<td>( p_k = -\nabla f(\bar{x}_k) / |\nabla f(\bar{x}_k)| )</td>
<td>Linear</td>
</tr>
<tr>
<td>Quasi-Newton</td>
<td>( p_k = -H_k^{-1} \nabla f(\bar{x}_k) )</td>
<td>Super-Linear</td>
</tr>
<tr>
<td>Newton</td>
<td>( p_k = -[\nabla^2 f(\bar{x}_k)]^{-1} \nabla f(\bar{x}_k) )</td>
<td>Quadratic</td>
</tr>
</tbody>
</table>

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

We will return to the selection of \( \bar{p}_k \), but let’s consider the computation
of the step length \( \alpha_k \)…

Peter Blomgren, ⟨blomgren.peter@gmail.com⟩
Line Search Methods: Step Length Selection

Given a descent direction \( \bar{p}_k \) we would like to find the global minimizer \( \alpha^*_k \) of

\[
\min_{\alpha > 0} f(\bar{x}_k + \alpha \bar{p}_k).
\]

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

— We want an \( \alpha_k \) so that we get a \textbf{substantial reduction} in the objective \( f \).

— We want to find \( \alpha_k \) \textbf{fast}.

In practice we perform an \textbf{inexact line search} — settling for an \( \alpha_k \) which gives \textbf{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.

Clearly, we need a stronger condition than $f(x_k + \alpha_k p_k) < f(x_k)$.
There are many ways to enforce reduction in the objective, e.g.

**Armijo Condition**

The Armijo Condition

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

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

To rule out unacceptably short steps, we additionally enforce

**Curvature Condition**

(Wolfe Condition #2)

The **Curvature Condition**

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

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

Together these two conditions are known as the **Wolfe conditions**.
The **Armijo Condition**

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

requires the reduction to be proportional to the step length \(\alpha\), as well as the directional derivative. **In practice** \(c_1\) is usually set to be quite small, e.g. \(\sim 10^{-4}\).
To rule out unacceptable short steps, the **curvature condition**

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

— it prevents us from stopping when more progress can be made by moving further (increasing \( \alpha \)). Typical values: \( c_{2N,QN}^2 = 0.9, \quad c_{2CG}^2 = 0.1 \).
Together, the Armijo and Curvature conditions constitute the Wolfe Conditions.
Together, the Armijo and Curvature conditions constitute the Wolfe Conditions.
The Strong Wolfe Conditions

A step length $\alpha$ may satisfy the **Wolfe Conditions**

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

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

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

further disallows values of

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

which are “too positive,” thus excluding points that are far from the stationary points of $\bar{p}_k^T \nabla f(\bar{x}_k + \alpha \bar{p}_k)$.
Are the Wolfe Conditions too Restrictive?

It can be shown (see NW$^2$nd pp.35–36) that there exist step lengths $\alpha$ 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 $\alpha$)**

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Let $\vec{p}_k$ be a descent direction at $\bar{x}_k$, and assume that $f$ is bounded below along the line \( \{ \bar{x}_k + \alpha \vec{p}_k : \alpha > 0 \} \). 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$^2$nd p.36.)
Algorithm: Backtracking Linesearch

[0] Find a descent direction $\bar{p}_k$
[1] Set $\alpha > 0$, $\rho \in (0,1)$, $c \in (0,1)$, set $\alpha = \bar{\alpha}$
[2] While $f(\bar{x}_k + \alpha \bar{p}_k) > f(\bar{x}_k) + c \alpha \bar{p}_k^T \nabla f(\bar{x}_k)$
[3] $\alpha = \rho \alpha$
[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 ($\bar{p}_k = \bar{p}_k^N$), where $\bar{\alpha} = 1$. It is less successful for quasi-Newton and CG-based approaches.

The value of the contraction factor $\rho$ can be allowed to vary at each iteration of the line search. (To be revisited)
NW2nd-3.1: Program the steepest descent and Newton algorithms using the backtracking line search. Use them to minimize the Rosenbrock function

\[ f(\bar{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 \]

Set the initial step length \( \alpha_0 = 1 \) and report the step length used by each method at each iteration. First try the initial point \( \bar{x}_0^T = [1.2, 1.2] \) and then the more difficult point \( \bar{x}_0^T = [-1.2, 1] \).

Note: The homework is due in Peter’s mailbox in GMCS-411 or in Peter’s office GMCS-587 (slide under the door if I’m not there).
Armijo condition, 18
Convergence
  Linear, 5
  Quadratic, 5
  Rate of, 4
  Superlinear, 5
Curvature condition, 19
Wolfe conditions, 19