Numerical Optimization

Lecture Notes #4
Unconstrained Optimization; 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 2018

Outline

Quick Recap: Last Time

Quick overview of optimization algorithms — two categories

1. **Line search** — Select a search direction, then optimize in that direction
2. **Trust region** — Build a local (simple) model of the objective, optimize the model in the local region where we trust it.

Convergence rates — linear / superlinear / quadratic.

More details on line search —

Search directions — Steepest descent (linear convergence)
Newton direction (quadratic convergence)

Enforcing sufficient decrease in the objective — **Wolfe conditions**.

Recall — Algorithm: Backtracking Linesearch

Algorithm: Backtracking Linesearch

```
[0] Find a descent direction p̄_k
[1] Set ᾱ > 0, ρ ∈ (0,1), c ∈ (0,1), set α = ᾱ
[2] While f(x̄_k + α p̄_k) > f(x̄_k) + c α p̄_k^T ∇f(x̄_k)
[3]     α = ρα
[4] End-While
[5] Set α_k = α
```

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 $\overline{\alpha} = \mathbf{1}$.

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

Example: Minimizing $f(\bar{\mathbf{x}}) = (x_1 + x_2^2)^2$

In the next few slides we illustrate our findings so far by minimizing $f(\bar{\mathbf{x}}) = (x_1 + x_2^2)^2$ (with $\bar{\mathbf{x}}_0 = [1, 1]$) using two algorithms:

— 1. Steepest Descent direction with Backtracking Linesearch.
— 2. Newton direction with Backtracking Linesearch.

One thing to notice is that the entire curve where $x_1 = -x_2^2$ gives $f(\bar{\mathbf{x}}) = 0$ — the minimum is not isolated, nor unique.

Recall:

$$\bar{\mathbf{p}}_k^{\mathrm{SD}} = -\frac{\nabla f(\bar{\mathbf{x}}_k)}{|\nabla f(\bar{\mathbf{x}}_k)|}, \qquad \bar{\mathbf{p}}_k^{N} = -\left[\nabla^2 f(\bar{\mathbf{x}}_k)\right]^{-1} \nabla f(\bar{\mathbf{x}}_k).$$

Steepest Descent on $f(\bar{\mathbf{x}}) = (x_1 + x_2^2)^2$, with $\bar{\mathbf{x}}_0 = [1, 1]$

$\bar{\mathbf{x}}^T$	$f(\bar{\mathbf{x}})$	$\bar{\mathbf{p}}^{\mathrm{SD}}$	α
$[1, 1]$	4.000000	$[-4.472136 \cdot 10^{-1}, -8.944272 \cdot 10^{-1}]$	1
$[0.552786, 0.105573]$	$3.180193 \cdot 10^{-1}$	$[-9.784275 \cdot 10^{-1}, -2.065907 \cdot 10^{-1}]$	1
$[-0.425641, -0.101018]$	$1.725874 \cdot 10^{-1}$	$[9.801951 \cdot 10^{-1}, -1.980344 \cdot 10^{-1}]$	$1/2$
$[0.064456, -0.200035]$	$1.091409 \cdot 10^{-2}$	$[-9.284542 \cdot 10^{-1}, 3.714468 \cdot 10^{-1}]$	$1/8$
$[-0.051600, -0.153604]$	$7.843386 \cdot 10^{-4}$	$[9.559089 \cdot 10^{-1}, -2.936633 \cdot 10^{-1}]$	$1/32$
$[-0.021728, -0.162781]$	$2.274880 \cdot 10^{-5}$	$[-9.508768 \cdot 10^{-1}, 3.095697 \cdot 10^{-1}]$	$1/128$
$[-0.029157, -0.160363]$	$1.183829 \cdot 10^{-5}$	$[9.522234 \cdot 10^{-1}, -3.054022 \cdot 10^{-1}]$	$1/256$
$[-0.025437, -0.161556]$	$4.395456 \cdot 10^{-7}$	$[-9.515610 \cdot 10^{-1}, 3.074601 \cdot 10^{-1}]$	$1/1024$
$[-0.026367, -0.161255]$	$1.319155 \cdot 10^{-7}$	$[9.517280 \cdot 10^{-1}, -3.069426 \cdot 10^{-1}]$	$1/2048$
$[-0.025902, -0.161405]$	$2.246036 \cdot 10^{-8}$	$[-9.516447 \cdot 10^{-1}, 3.072010 \cdot 10^{-1}]$	$1/4096$
$[-0.026134, -0.161330]$	$1.137904 \cdot 10^{-8}$	$[9.516864 \cdot 10^{-1}, -3.070717 \cdot 10^{-1}]$	$1/8192$

Table: Steepest Descent with Backtracking line search applied to the problem $f(\bar{\mathbf{x}}) = (x_1 + x_2^2)^2$, starting $\bar{\mathbf{x}}_0 = [1, 1]$.

Steepest Descent with Backtracking Linesearch

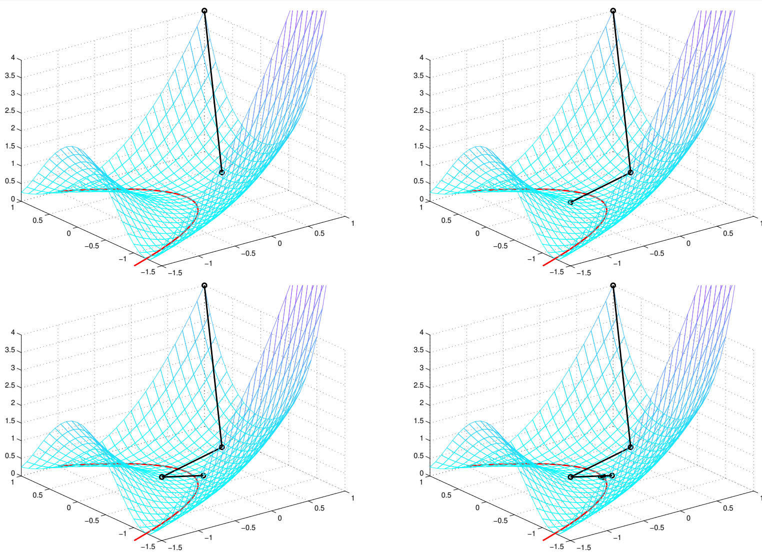

Figure: The iterations #1, #2, #3, and #9 (convergence to 10^{-8}).

Steepest Descent with Backtracking Linesearch

Linear Convergence

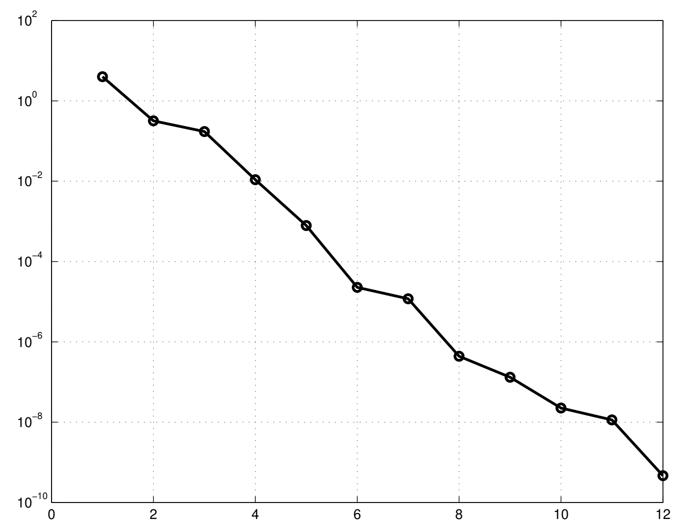

Figure: The size of the objective as a function of the iteration number.

Newton on $f(\bar{\mathbf{x}}) = (x_1 + x_2^2)^2$, with $\bar{\mathbf{x}}_0 = [1, 1]$

Oops! Non-local Minimum

$\bar{\mathbf{x}}^T$	$f(\bar{\mathbf{x}})$	p^N	α
$[1, 1]$	4.000000	$[-2.000000, 0.000000]$	1
$[-1, 1]$	0	—	—

Table: Newton direction with Backtracking line search applied to the problem $f(\bar{\mathbf{x}}) = (x_1 + x_2^2)^2$, starting $\bar{\mathbf{x}}_0 = [1, 1]$.

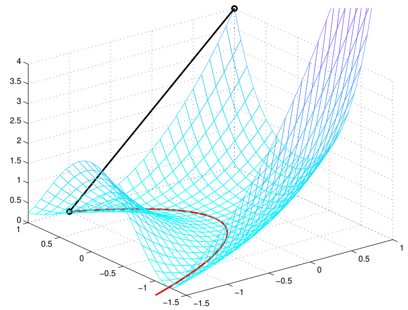

Figure: The only iteration; we achieve convergence right away. ($\Rightarrow$ The problem is too easy...)

Example #2: Minimizing $f(\bar{\mathbf{x}}) = (x_1 + x_2^2)^2 + 0.5(x_1^2 + x_2^2)$

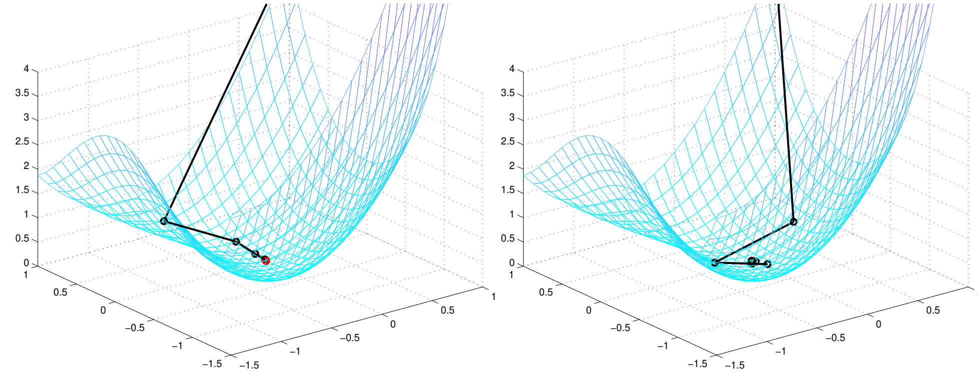

Figure: Comparison of Newton direction (left) and Steepest Descent direction (right) with backtracking line search applied to the objective $f(\bar{\mathbf{x}}) = (x_1 + x_2^2)^2 + 0.5(x_1^2 + x_2^2)$ with $\bar{\mathbf{x}}_0^T = [1, 1]$. In this case, the minimum is unique $\bar{\mathbf{x}}^* = [0, 0]$. The Newton algorithm requires 6 iterations to reach the minimum up to desired accuracy, and the Steepest Descent algorithm requires 16 iterations.

Example: Minimizing $f(\bar{\mathbf{x}}) = (x_1 + x_2^2)^2 + 0.5(x_1^2 + x_2^2)$

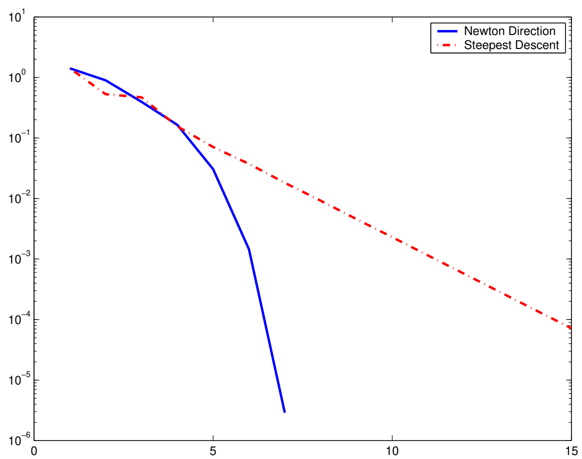

Figure: Convergence comparison of Newton direction (blue/solid) and Steepest Descent direction (red/dash-dotted) with backtracking line search applied to the objective $f(\bar{\mathbf{x}}) = (x_1 + x_2^2)^2 + 0.5(x_1^2 + x_2^2)$ with $\bar{\mathbf{x}}_0^T = [1, 1]$. We see that the Steepest Descent algorithm is linearly convergent and that the Newton algorithm is significantly faster (quadratically convergent).

Convergence of Line Search Methods

$\cos\theta_k$

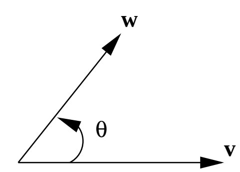

Figure: $\bar{\mathbf{v}}^T\bar{\mathbf{w}} = \cos\theta\,\|\bar{\mathbf{v}}\| \cdot \|\bar{\mathbf{w}}\|$.

All our algorithms rely on the directional derivative $\bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}_k)$.

One of the key properties in expressing requirements on the search direction is the **angle** θ_k.

The expression for the cos of this angle is given by

$$\cos\theta_k = \frac{\bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}_k)}{\|\bar{\mathbf{p}}_k^T\|\,\|\nabla f(\bar{\mathbf{x}}_k)\|}.$$

Convergence of Line Search Methods

Theorem (Zoutendijk's Theorem)

Consider any iteration of the form $\bar{\mathbf{x}}_{k+1} = \bar{\mathbf{x}}_k + \alpha_k \bar{\mathbf{p}}_k$, *where* $\bar{\mathbf{p}}_k$ *is a descent direction and* α_k *satisfies the Wolfe conditions*

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

Suppose that f *is bounded from below in* $\mathbb{R}^n$ *and that* f *is continuously differentiable in an open set* N *containing the level set* $\mathcal{L} = \{x \in \mathbb{R}^n : f(\bar{\mathbf{x}}) \leq f(\bar{\mathbf{x}}_0)\}$ *where* $\bar{\mathbf{x}}_0$ *is the starting point of the iteration. Assume also that the gradient* ∇f *is* **Lipschitz continuous** *on* N, *i.e. there exists a constant* L *such that*

$$\|\nabla f(\bar{\mathbf{x}}) - \nabla f(\bar{\mathbf{y}})\| \leq L \|\bar{\mathbf{x}} - \bar{\mathbf{y}}\|, \quad \forall \bar{\mathbf{x}}, \bar{\mathbf{y}} \in \mathcal{L}$$

Then

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f(\bar{\mathbf{x}}_k)\|^2 < \infty.$$

Convergence of Line Search Methods

This result seems, at first (and probably second too!) glance, to be quite technical and obscure...

However, the **Zoutendijk condition**

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f(\bar{\mathbf{x}}_k)\|^2 < \infty$$

implies that

$$\lim_{k \to \infty} \cos^2 \theta_k \|\nabla f(\bar{\mathbf{x}}_k)\|^2 \to 0$$

Now, **if** we have a method for choosing the search directions $\bar{\mathbf{p}}_k$ so that $\cos \theta_k \geq \delta > 0$ for all k, **then** this shows that $\|\nabla f(\bar{\mathbf{x}}_k)\| \to 0$.

The Zoutendijk Condition: Steepest Descent

For the steepest descent direction we have $\cos \theta_k = -1$. Hence, if we use a line search algorithm which satisfies the Wolfe conditions, it will always converge to a stationary point (under the conditions of the theorem — f bounded below, and ∇f Lipschitz continuous).

This means that steepest descent is **globally convergent** in the sense

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

We cannot guarantee convergence to a minimum, only to a stationary point.

In order to guarantee convergence to a minimum, more conditions (for example on the Hessian) are required.

The Zoutendijk Condition: Newton Direction

It can be shown that Newton methods are also globally convergent in this sense, under these additional conditions:

The Hessian must be **positive definite**, and the **condition numbers** must be uniformly bounded, *i.e.*

$$\left\| \left[\nabla^2 f(\bar{\mathbf{x}}_k)\right] \right\| \cdot \left\| \left[\nabla^2 f(\bar{\mathbf{x}}_k)\right]^{-1} \right\| \leq M, \quad \text{for some positive } M.$$

That is, the ratio of the largest and smallest eigenvalues, $\lambda_k^{\max} / \lambda_k^{\min}$ must remain bounded. (Think of λ_k^j as the curvature in the eigen-directions.)

The proof for the steepest descent direction is given in (NW^{1st} pp.43–44, NW^{2nd} pp.38–39) and the key part to the proof for the Newton direction is exercise #3.5 (NW^{1st} p.62, NW^{2nd} p.63)

Notes & Looking Forward

We note that the Zoutendijk condition gives us global convergence, *i.e.* a guarantee that we will arrive at a stationary point.

It does, however, not say anything about the **rate of convergence**.

Next we will look (in a less hand-waving way) at the **local convergence rates** (*i.e.* the convergence "speed") of Steepest descent, Newton, and Quasi-Newton methods.

Index