THE UNIVERSITY OF CHICAGO

Section 3: Line Search Methods

Mihai Anitescu STAT 310

Reference: Chapter 3 in Nocedal and Wright.

Line Search Methods Idea:

- At the current point x_k find a "Newton-like" direction d_k
- Along that direction d_k do 1-dimensional minimization (simpler than over whole space)

$$x_{k+1} \approx \arg\min_{\alpha} f(x_k + \alpha d_k)$$

- Because the line search always decreases f, we will have an accumulation point (cannot diverge if bounded below) – unlike Newton proper

Descent Principle

- Descent Principle: Carry Out a one-Dimensional Search Along a Line where I will decrease the function. The descent condition is

$$\nabla f(x_k)^T p_k < 0$$

- If this happens, there exists an alpha such that.

$$f\left(x_k + \alpha p_k\right) < f\left(x_k\right)$$

- So I will keep making progress.
- Typical choice $B_k p_k = -\nabla f(x_k); \quad B_k \succ 0$
- B_k may sometimes be the Hessian, but in general Newton may need to be modified

3.1 Choosing Step Length. Step Conditions

- Solving the exact minimization problem may be too expensive, even in 1D:

$$\arg\min_{\alpha} f(x_k + \alpha d_k)$$

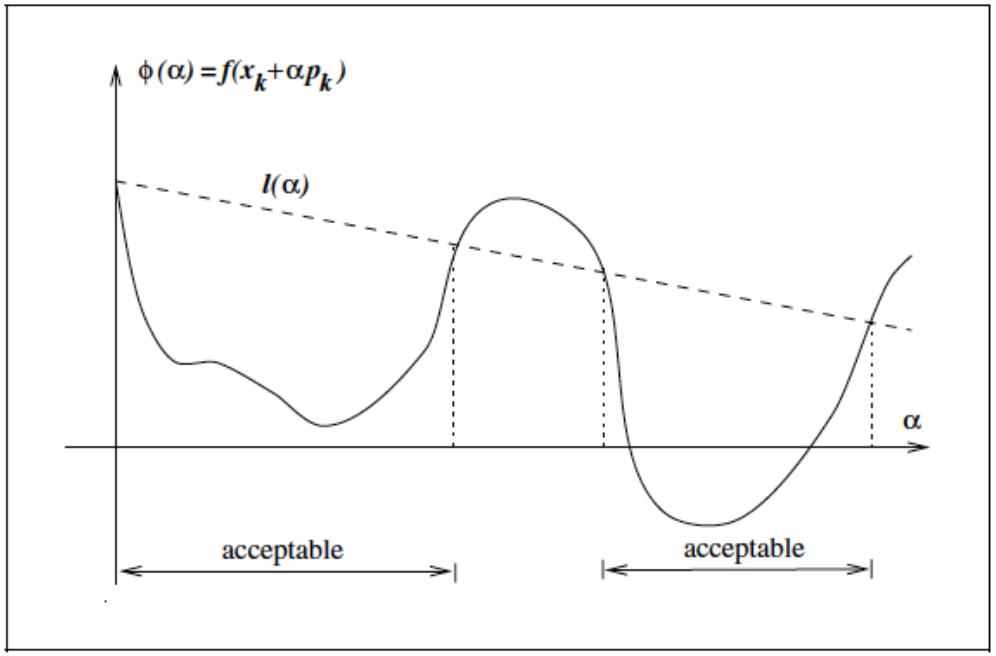
- The condition that we solve this problem exactly is replaced by step conditions that
  - Ensure global convergence (in the sense described in objectives)
  - Ensure that the (~) Newton step is accepted – if it provides sufficient decrease.
- Accepting only descent may not be sufficient for global convergence: consider f(x)=(x-1)^2-1, and

$$x_k = -\frac{1}{k}; \quad f(x_k) = \frac{2}{k} + \frac{1}{k^2} < f(x_{k+1}) \qquad x_k \to 0 \neq 1 = x^*$$

The Wolfe Conditions—sufficient descent.

- Objective 1: Produce *sufficient decrease ('don't gc VERY far ...')*

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k$$



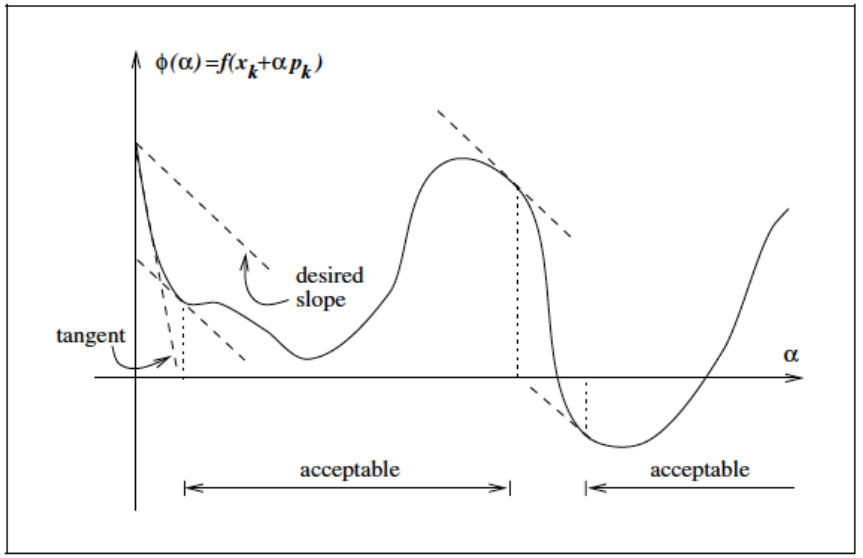
- In practice, c_1=1e-4 or so (small!) – do not want to eliminate Newton steps if ill-conditioned.
- However, by itself does not resolve the problem of arbitrary small steps.

The Wolfe Conditions—Curvature Condition

- Objective 2: sufficiently large step ("… but far enough")
- Insight: at the minimum, the directional derivative is zero.
- Curvature conditions:

$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k,$$

- Requ $0 < c_1 < c_2 < 1$



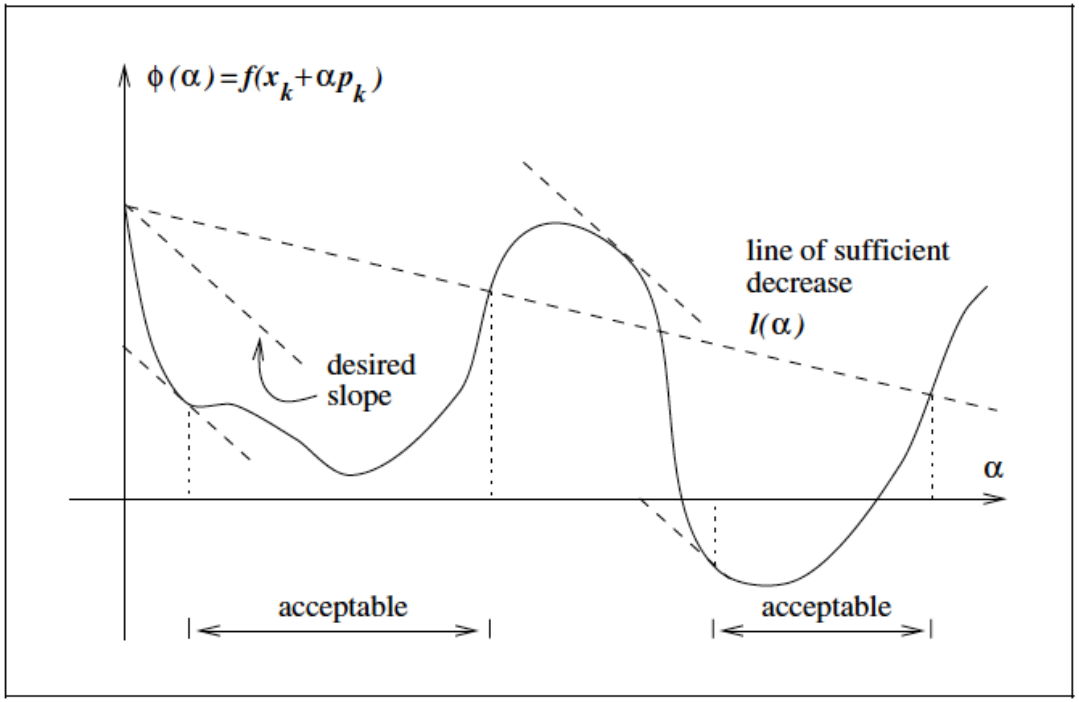
- Typical choices: c_2=0.9 for Newton and Quasi-Newton, and 0.1 for Conjugated Gradient.

Wolfe Conditions

- In summary:

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f_k^T p_k,$$
$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k,$$

- Graphics:



Wolfe Conditions:Theory

Lemma 3.1.

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ *is continuously differentiable. Let* p_k *be a descent direction at* x_k*, and assume that* f *is bounded below along the ray* $\{x_k + \alpha p_k | \alpha > 0\}$*. Then if* $0 < c_1 < c_2 < 1$*, there exist intervals of step lengths satisfying the Wolfe conditions (3.6) and the strong Wolfe conditions (3.7).*

- Is bounded below a reasonable assumption?
- Proof.

Backtracking(Armijo)

- Idea: Accept the *largest* stepsize that satisfies the *descent condition* (this dispenses with the curvature condition in Wolfe or lower bound in Goldstein).

Algorithm 3.1 (Backtracking Line Search).
Choose $\bar{\alpha} > 0, \rho \in (0, 1), c \in (0, 1)$; Set $\alpha \leftarrow \bar{\alpha}$;
repeat until $f(x_k + \alpha p_k) \leq f(x_k) + c\alpha \nabla f_k^T p_k$
$\alpha \leftarrow \rho\alpha$;
end (repeat)
Terminate with $\alpha_k = \alpha$.

- For Newton and quasi-Newton want $\bar{\alpha} = 1$ (to allow for quadratic convergence).
- For quasi-Newton and CG less appropriate.
- Large algorithmic variability: e.g. can choose

$$\rho \in [\rho_{lo}, \rho_{hi}], \quad 0 < \rho_{lo} < \rho_{hi} < 1$$

3.2 Convergence Theory. Theorem (Zoutendijk)

Theorem 3.2.

Consider any iteration of the form (3.1), where p_k is a descent direction and α_k satisfies the Wolfe conditions (3.6). Suppose that f is bounded below in $\mathbb{R}^n$ and that f is continuously differentiable in an open set $\mathcal{N}$ containing the level set $\mathcal{L} \stackrel{\text{def}}{=} \{x : f(x) \le f(x_0)\}$, 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

$$\|\nabla f(x) - \nabla f(\tilde{x})\| \le L\|x - \tilde{x}\|, \qquad \text{for all } x,\ \tilde{x} \in \mathcal{N}.$$

Then:
$$\sum_{k>0} \cos^2\theta_k \, \|\nabla f_k\|^2 < \infty.$$

Here:
$$\cos\theta_k = \frac{-\nabla f_k^T p_k}{\|\nabla f_k\| \, \|p_k\|}$$

Consequences of Zoutendijk Theorem

- Consequence: if the Wolfe conditions hold, then we must have

$$\cos^2 \theta_k \|\nabla f_k\|^2 \to 0.$$

- In turn, if the angle the descent direction is bounded away from 90 degrees, that is,

$$\cos \theta_k \geq \delta > 0, \quad \text{for all } k.$$

- We are guaranteed to converge to a first-order stationary point, our first global convergence objective.

$$\lim_{k \to \infty} \|\nabla f_k\| = 0.$$

How might we ensure an angle bounded away from 90 degrees?

- Consider a Newton-Like method

$$x_{k+1} = x_k + \alpha_k p_k,$$

$$p_k = -B_k^{-1} \nabla f_k,$$

- Assume now that the matrix B_k is bounded and uniformly positive definite, that is

$$\|B_k\| \, \|B_k^{-1}\| \leq M, \quad \text{for all } k.$$

- It then follows that $\cos\theta_k \geq 1/M$ and, from Zoutendijk,

$$\lim_{k \to \infty} \|\nabla f_k\| = 0.$$

3.3Rate of Convergence.

Convergence rate of steepest descent

- If the objective function is quadratic $f(x) = \frac{1}{2}x^T Qx - b^T x$.

Its gradient is simply $\nabla f(x) = Qx - b$

- We get for the line search

$$\alpha_k = \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k}. \qquad x_{k+1} = x_k - \left(\frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k} \right) \nabla f_k$$

- In the weighted norm, $\|x\|_Q^2 = x^T Qx$ we have the identity:

$$\|x_{k+1} - x^*\|_Q^2 = \boxed{\left\{ 1 - \frac{\left(\nabla f_k^T \nabla f_k\right)^2}{\left(\nabla f_k^T Q \nabla f_k\right)\left(\nabla f_k^T Q^{-1} \nabla f_k\right)} \right\}} \|x_k - x^*\|_Q^2$$

- By analyzing the boxed expression, we get

Result

Theorem 3.3.

When the steepest descent method with exact line searches (3.26) is applied to the strongly convex quadratic function (3.24), the error norm (3.27) satisfies

$$\|x_{k+1} - x^*\|_Q^2 \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 \|x_k - x^*\|_Q^2,$$

where $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ *are the eigenvalues of* Q.

- Because the ratio is less than 1, this indicates a linear rate of convergence.
- The ratio gets worse as we matrix Q is more ill-conditioned, i.e. it gets closer to one.
- Recall the zig-zag pattern I mentioned in Section 2.
- Nonlinear case, very similar result.

Newton's Method

Theorem 3.5.

Suppose that f is twice differentiable and that the Hessian $\nabla^2 f(x)$ is Lipschitz continuous (see (A.42)) in a neighborhood of a solution x^* at which the sufficient conditions (Theorem 2.4) are satisfied. Consider the iteration $x_{k+1} = x_k + p_k$, where p_k is given by (3.30). Then

(i) *if the starting point x_0 is sufficiently close to x^*, the sequence of iterates converges to x^*;*

(ii) *the rate of convergence of $\{x_k\}$ is quadratic; and*

(iii) *the sequence of gradient norms $\{\|\nabla f_k\|\}$ converges quadratically to zero.*

- (Full proof)
- Here, Lipschitz continuity is weaker than three times continuous differentiability:

$$\|\nabla^2_{xx} f(x_2) - \nabla^2_{xx} f(x_1)\| \leq L\|x_2 - x_1\|$$

Approximate Newton Directions

Theorem 3.6.

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. Consider the iteration $x_{k+1} = x_k + \alpha_k p_k$, where p_k is a descent direction and α_k satisfies the Wolfe conditions (3.6) with $c_1 \leq 1/2$. If the sequence $\{x_k\}$ converges to a point x^* such that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite, and if the search direction satisfies

$$\lim_{k \to \infty} \frac{\|\nabla f_k + \nabla^2 f_k p_k\|}{\|p_k\|} = 0,$$

then

(i) *the step length $\alpha_k = 1$ is admissible for all k greater than a certain index k_0; and*

(ii) *if $\alpha_k = 1$ for all $k > k_0$, $\{x_k\}$ converges to x^* superlinearly.*

- Note: This is satisfied by Newton's method, but it allows us to go for only approximate directions, as built by Quasi-Newton methods.

Quasi-Newton Methods

- For Quasi-Newton methods we have that $-\nabla f(x_k) = B_k p_k$ and thus the condition becomes:

$$\lim_{k \to \infty} \frac{\|(B_k - \nabla^2 f(x^*))p_k\|}{\|p_k\|} = 0$$

- Turns out this conditions *is necessary and sufficient* for superlinear convergence.

Theorem 3.7.

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. Consider the iteration $x_{k+1} = x_k + p_k$ (that is, the step length α_k is uniformly 1) and that p_k is given by (3.34). Let us assume also that $\{x_k\}$ converges to a point x^* such that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite. Then $\{x_k\}$ converges superlinearly if and only if (3.36) holds.

3.4 NEWTON's METHOD WITH HESSIAN MODIFICATION

- If the problem is not convex, far away from the solution, Newton's direction is not necessarily a descent direction:

$$\nabla^2_{xx} f(x_k) p_k^N = -\nabla f(x_k) \Rightarrow {p_k^N}^T \nabla^2_{xx} f(x_k) p_k^N = -\nabla f(x_k)^T p_k$$

Thus, if the Hessian is indefinite it is possible the right side is >0.

- So, to use something like Newton's method, we need to modify the matrix.
  - Compute eigenvalues, make it positive definite (expensive, even if H sparse!)
  - Try to do Cholesky, if fail, keep adding multiples of identity (algo 3.3). Would throw it all away until it succeeds.
  - Modified Cholesky – Need to implement it yourself.
  - Modified Bunch Parlett --

General framework for Modified Newton Method

Algorithm 3.2 (Line Search Newton with Modification).

Given initial point x_0;

for $k = 0, 1, 2, \ldots$

Factorize the matrix $B_k = \nabla^2 f(x_k) + E_k$, where $E_k = 0$ if $\nabla^2 f(x_k)$ is sufficiently positive definite; otherwise, E_k is chosen to ensure that B_k is sufficiently positive definite;

Solve $B_k p_k = -\nabla f(x_k)$;

Set $x_{k+1} \leftarrow x_k + \alpha_k p_k$, where α_k satisfies the Wolfe, Goldstein, or Armijo backtracking conditions;

end

LDL factorization WITH permutation

- Bunch-Parlett, or Bunch-Kaufmann:

$$PAP^T = LBL^T$$

- For any A symmetric, there exists the lower triangular matrix L, the permutation matrix P and the block diagonal B made of 1 by 1 and 2 by 2 blocks such that the above holds.
- B has the same signature as A (number of positive and negative eigenvalues) – Sylvester's law of inertia.
- The amount of computation is comparable to Cholesky.

Modified Bunch Parlett

- Modify the matrix B: $L(B+F)L^T$
- Where the modification F satisfies:

$$F = Q\,\mathrm{diag}(\tau_i)\,Q^T, \qquad \tau_i = \begin{cases} 0, & \lambda_i \geq \delta, \\ \delta - \lambda_i, & \lambda_i < \delta, \end{cases} \qquad i = 1, 2, \ldots, n,$$

- Thus B+F must be positive definite, and then, so is $L(B+F)L^T$
- Therefore the modification E will satisfy:

$$P(A+E)P^T = L(B+F)L^T, \qquad \text{where } E = P^T L F L^T P$$

3.5 Some musings: Step size selection algorithm.

- Objective: reduce the number of function evaluations.
- Backtracking (Armijo) is easy to code (and not too bad practically, as it can be warm-started),
- It can be modified into Wolfe if we add something like bisection.
- It can be improved by creating a one dimensional function model (interpolation)
- Professional line search algorithms are very sophisticated.