Lecture 5: CS395T Numerical Optimization for Graphics and AI — Line Search

Qixing Huang
The University of Texas at Austin
huangqx@cs.utexas.edu

1 Disclaimer

This note is adapted from Section 3 of

- *Numerical Optimization* by Jorge Nocedal and Stephen J. Wright. Springer series in operations research and financial engineering. Springer, New York, NY, 2. ed. edition, (2006)

2 Introduction

Each iteration of a line search method computes a search direction $\boldsymbol{p}_k$ and then decides how far to move along that direction. The iteration is given by

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

where the positive scalar α_k is called the step length. The success of a line search method depends on effective choices of both the direction $\boldsymbol{p}_k$ and the step length α_k. Most line search algorithms require $\boldsymbol{p}_k$ to be a descent direction one for which $\boldsymbol{p}_k^T \nabla f(\boldsymbol{x}_k) < 0$—because this property guarantees that the function f can be reduced along this direction, as discussed in the previous chapter. Moreover, the search direction often has the form

$$\boldsymbol{p}_k = -B_k^{-1} \nabla f(\boldsymbol{x}_k), \tag{2}$$

where B_k is a symmetric and non-singular matrix. In the steepest descent method B_k is simply the identity matrix I , while in Newton's method B_k is the exact Hessian $\nabla^2 f(\boldsymbol{x}_k)$. In quasi-Newton methods, B_k is an approximation to the Hessian that is updated at every iteration by means of a low-rank formula. When $\boldsymbol{p}_k$ is defined by (2) and B_k is positive definite, we have

$$\boldsymbol{p}_k^T \nabla f(\boldsymbol{x}_k) = -\nabla f(\boldsymbol{x}_k)^T B_k^{-1} \nabla f(\boldsymbol{x}_k) < 0,$$

and therefore $\boldsymbol{p}_k$ is a descent direction. In the next few lectures we study how to choose the matrix B_k, or more generally, how to compute the search direction. We now give careful consideration to the choice of the step-length parameter α_k.

3 Step Length

In computing the step length α_k, we face a tradeoff. We would like to choose α_k to give a substantial reduction of f, but at the same time, we do not want to spend too much time making the choice. The ideal choice would be the global minimizer of

$$\phi(\alpha) = f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k), \alpha > 0, \tag{3}$$

but in general, it is too expensive to identify this value. To find even a local minimizer of ϕ to moderate precision generally requires too many evaluations of the objective function f and possibly the gradient ∇f. More practical strategies perform an inexact line search to identify a step length that achieves adequate reductions in f at minimal cost. Typical line search algorithms try out a sequence of candidate values for α, stopping to accept one of these values when certain conditions are satisfied. The line search is done in two stages: A bracketing phase finds an interval containing desirable step lengths, and a bisection or interpolation phase computes a good step length within this interval.

How to terminate a line-search is critical. We now discuss various termination conditions for the line search algorithm and show that effective step lengths need not lie near minimizers of the univariate function $\phi(\alpha)$ defined above. A simple condition we could impose on α_k is that it provide a reduction in f , i.e., $f(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k) < f(\boldsymbol{x}_k)$. However, this may not be appropriate. For example, consider minimizing $f(x) = (x+1)^2$. We can let $x_k = \frac{5}{k}$. It is clear that the value of the objective function always goes down. However, it does not go the minimizer. The difficulty is that we do not have sufficient reduction in f, a concept we discuss next.

3.1 The Wolfe Conditions

A popular inexact line search condition stipulates that α_k should first of all give sufficient decrease in the objective function f, as measured by the following inequality:

$$f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k) \leq f(\boldsymbol{x}_k) + c_1 \alpha \nabla f(\boldsymbol{x}_k)^T \boldsymbol{p}_k, \tag{4}$$

for some constant $c_1 \in (0, 1)$. In other words, if you move far, then you should have more reduction. This implicitly favors small step-sizes. In practice, c_1 is chosen to be small, i.e., $c_1 = 10^{-4}$.

The sufficient decrease condition is not enough by itself to ensure that the algorithm makes reasonable progress. To rule out unacceptably short steps we introduce a second requirement, called the curvature condition, which requires α_k to satisfy

$$\nabla f(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k)^T \boldsymbol{p}_k \geq c_2 \nabla f(\boldsymbol{x}_k)^T \boldsymbol{p}_k, \tag{5}$$

for some constant $c_2 \in (c_1, 1)$, where c_1 is the constant from (4). Note that the left-hand-side is simply the derivative $\phi'(\alpha_k)$, so the curvature condition ensures that the slope of $\phi(\alpha_k)$ is greater than c_2 times the gradient $\phi'(0)$. This makes sense because if the slope $\phi(\alpha)$ is strongly negative, we have an indication that we can reduce f significantly by moving further along the chosen direction. On the other hand, if the slope is only slightly negative or even positive, it is a sign that we cannot expect much more decrease in f in this direction, so it might make sense to terminate the line search. Typical values of c_2 are 0.9 when the search direction $\boldsymbol{p}_k$ is chosen by a Newton or quasi-Newton method, and 0.1 when p k is obtained first order methods.

The sufficient decrease and curvature conditions are known collectively as the *Wolfe conditions*. We restate them here for future reference:

$$f(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k) \leq f(\boldsymbol{x}_k) + c_1 \alpha_k \nabla f(\boldsymbol{x}_k)^T \boldsymbol{p}_k, \tag{6}$$

$$\nabla f(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k)^T \boldsymbol{p}_k \geq c_2 \nabla f(\boldsymbol{x}_k)^T \boldsymbol{p}_k, \tag{7}$$

with $0 < c_1 < c_2 < 1$.

Lemma 3.1. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Let $\boldsymbol{p}_k$ be a decent direction at $\boldsymbol{x}_k$, and assume that f is bounded below along the ray $\{\boldsymbol{x}_k + \alpha \boldsymbol{p}_k | \alpha > 0\}$. Then if $0 < c_1 < c_2 < 1$, there exist intervals of step lengths satisfying the Wolfe conditions.*

The proof is straight-forward using mean-value theorem.

3.2 Sufficient Decrease and Backtracking

We have mentioned that the sufficient decrease condition alone is not sufficient to ensure that the algorithm makes reasonable progress along the given search direction. However, if the line search algorithm chooses its candidate step lengths appropriately, by using a so-called backtracking approach, we can dispense with the extra condition use just the sufficient decrease condition to terminate the line search procedure. In its most basic form, backtracking proceeds as follows:

- Choose $\overline{\alpha} > 0, \rho, c \in (0,1)$; set $\alpha \leftarrow \overline{\alpha}$;
- **repeat** until $f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k) \leq f(\boldsymbol{x}_k) + c\alpha \nabla f(\boldsymbol{x}_k)^T \boldsymbol{p}_k$.
- $\alpha \leftarrow \rho\alpha$;
- end.

4 Convergence of Line Search Methods

To obtain global convergence, we must not only have well-chosen step lengths but also well-chosen search directions $\boldsymbol{p}_k$. We discuss requirements on the search direction in this section, focusing on one key property: the angle θ_k between $\boldsymbol{p}_k$ and the steepest descent direction $-\nabla f(\boldsymbol{x}_k)$, defined by

$$\cos(\theta_k) = -\frac{\nabla f(\boldsymbol{x}_k)^T \boldsymbol{p}_k}{\|\nabla f(\boldsymbol{x}_k)\| \|\boldsymbol{p}_k\|}. \tag{8}$$

The following theorem, due to Zoutendijk, has far-reaching consequences. It shows, for example, that the steepest descent method is globally convergent. For other algorithms it describes how far $\boldsymbol{p}_k$ can deviate from the steepest descent direction and still give rise to a globally convergent iteration. Various line search termination conditions can be used to establish this result, but for concreteness we will consider only the Wolfe conditions.

Theorem 4.1. *Consider any iteration of the form (1), where $\boldsymbol{p}_k$ is a descent direction and α_k satisfies the Wolfe conditions. 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 $\mathcal{L} := \{\boldsymbol{x} : f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$, where $\boldsymbol{x}_0$ 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*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\bar{\boldsymbol{x}})\| \leq L\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|, \quad \forall \boldsymbol{x}, \bar{\boldsymbol{x}} \in \mathcal{N}. \tag{9}$$

Then

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

Proof: From the line search conditions, we have that

$$(\nabla f(\boldsymbol{x}_{k+1}) - \nabla f(\boldsymbol{x}_k))^T \boldsymbol{p}_k \geq (c_2 - 1) \nabla f(\boldsymbol{x}_k)^T \boldsymbol{p}_k,$$

while the Lipschitz condition implies that

$$(\nabla f_{k+1} - \nabla f_k)^T \boldsymbol{p}_k \leq \alpha_k L \|\boldsymbol{p}_k\|^2.$$

By combing these two relations, we obtain

$$\alpha_k \geq \frac{c_2 - 1}{L} \frac{\nabla f(\boldsymbol{x}_k)^T \boldsymbol{p}_k}{\|\boldsymbol{p}_k\|^2}.$$

By substituting this inequality into the first Wolfe condition, we obtain

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - c_1 \frac{1 - c_2}{L} \frac{(\nabla f(\boldsymbol{x}_k)^T \boldsymbol{p}_k)^2}{\|\boldsymbol{p}_k\|^2},$$

or in other words,

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - c\cos^2(\theta_k)\|\nabla f(\boldsymbol{x}_k)\|^2,$$

where $c = c_1(1-c_2)/L$. We can conclude the proof by summing this expression over all indices less than or equal to k, we obtain

$$f_{k+1} \leq f_0 - c\sum_{j=0}^{k}\cos^2(\theta_j)\|\nabla f(\boldsymbol{x}_j)\|^2.$$

□

Steepest decent: $\lim_{k\to\infty}\|\nabla f(\boldsymbol{x}_k)\| = 0$.

Newton method:

$$\cos(\theta_k) \geq \frac{1}{M},$$

where $\|B_k\|\|B_k^{-1}\| \leq M$.

4.1 Convergence Rate

Convergence rate of steepest descent: Let us suppose that

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x},$$

where Q is symmetric and positive definite. The gradient is given by $\nabla f(\boldsymbol{x}) = Q\boldsymbol{x} - \boldsymbol{b}$, and the minimizer $\boldsymbol{x}^\star$ is the unique solution of the linear system $Q\boldsymbol{x} = \boldsymbol{b}$.

Let us compute the step length α_k that minimizes $f(\boldsymbol{x}_k - \alpha\nabla f(\boldsymbol{x}_k)$. By differentiating

$$f(\boldsymbol{x}_k - \alpha\boldsymbol{g}_k) = \frac{1}{2}(\boldsymbol{x}_k - \alpha\boldsymbol{g}_k)^T Q(\boldsymbol{x}_k - \alpha\boldsymbol{g}_k) - \boldsymbol{b}^T(\boldsymbol{x}_k - \alpha\boldsymbol{g}_k)$$

with respect to α, we obtain

$$\alpha = \frac{\nabla f(\boldsymbol{x}_k)^T\nabla f(\boldsymbol{x}_k)}{\nabla f(\boldsymbol{x}_k)^T Q\nabla f(\boldsymbol{x}_k)}.$$

If we use this exact minimizer α_k, the steepest descent iteration is given by

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \Big(\frac{\nabla f(\boldsymbol{x}_k)^T\nabla f(\boldsymbol{x}_k)}{\nabla f(\boldsymbol{x}_k)^T Q\nabla f(\boldsymbol{x}_k)}\Big)\nabla f(\boldsymbol{x}_k).$$

Theorem 4.2. *When the steepest descent method with exact line searches is applied, the error norm satisfies*

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\|_Q^2 \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 \|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_Q^2 \tag{10}$$

Note that for this function of interest,

$$\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^\star\|_Q^2 = (\boldsymbol{x} - \boldsymbol{x}^\star)^T Q(\boldsymbol{x} - \boldsymbol{x}^\star) = f(\boldsymbol{x}) - f(\boldsymbol{x}^\star).$$

The rate of convergence behavior of the steepest descent method is essentially the same on general nonlinear objective functions. In the following result we assume that the step length is the global minimizer along the search direction.

Theorem 4.3. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, and that the iterates generated by the steepest descent method with exact line searches converge to a point $\boldsymbol{x}^\star$ where the Hessian matrix $\nabla^2 f(\boldsymbol{x}^\star)$ is positive definite. Then*

$$(f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star)) \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 (f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star)). \tag{11}$$

Newton method:

Theorem 4.4. *Suppose that f is twice differentiable and that the Hessian $\nabla^2 f(\boldsymbol{x})$ is Lipschitz continuous in a neighborhood of a solution $\boldsymbol{x}^\star$ at which the sufficient conditions are satisfied. Consider the iteration $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{p}_k$, where $\boldsymbol{p}_k = -\big(\nabla^2 f(\boldsymbol{x}_k)\big)^{-1} \nabla f(\boldsymbol{x}_k)$. Then*

1. *if the starting point $\boldsymbol{x}_0$ is sufficiently close to $\boldsymbol{x}^\star$, the sequence of iterates converge to $\boldsymbol{x}^\star$;*
2. *the rate of convergence of $\boldsymbol{x}_k$ is quadratic; and*
3. *the sequence of gradient norms $\|\nabla f(\boldsymbol{x}_k)\|$ converges quadratically to zero.*