Lecture 20: Line Search Procedures; Newton's Method with Hessian Modification

Yudong Chen

1 Towards global convergence of Newton's method

Last time we consider the basic Newton's (BN) method $x_{k+1} = x_k - \alpha_k \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$. When f is strongly convex and has Lipschitz continuous Hessian, we show that BN achieves local quadratic convergence.

Global convergence does not hold in general for BN even if f is strongly convex, smooth and has Lipschitz Hessians, because the stepsize $\alpha_k = 1$ used BN could be too large. One solution (under these assumptions on f) is to consider a damped version of Newton's method:

$$x_{k+1} = x_k - \alpha_k \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k),$$

where the stepsize α_k is determined using a line search procedure.

We have discussed some line search procedures, including exact line search and backtracking line search, in the context of gradient descent (Lecture 7–8). In this lecture, we discuss more general line search procedures. They play an important role in Newton's methods and other second-order methods (such as quasi-Newton).

In f is nonconvex, the Hessian $\nabla^2 f(x)$ may be indefinite or singular. Further modification to BN is needed to ensure global convergence.

2 Line search procedures

Consider the general descent method:

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

where α_k is stepsize and p_k is a search direction, meaning that $\langle p_k, \nabla f(x_k) \rangle < 0$.

Ideally, we would like to choose α_k to minimize

$$\phi(\alpha) := f(x_k + \alpha p_k)$$

(i.e., exact line search). Finding the exact minimizer is often impractical (and unnecessary). Instead, we settle for a "good enough" α_k that satisfies certain conditions.

2.1 The Wolfe conditions

These conditions are most frequently used for nonlinear CG and quasi-Newton methods.

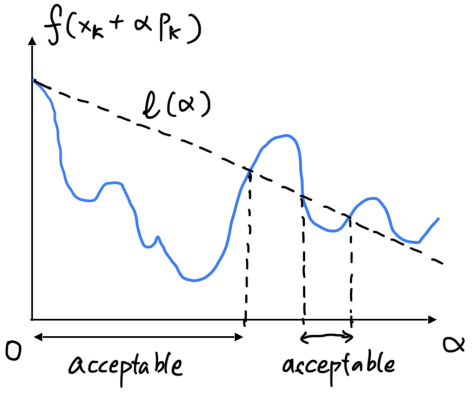
2.1.1 (Weak) Wolfe conditions

Let c_1, c_2 be two numbers satisfying $0 < c_1 < c_2 < 1$ (typically $c_1 = 10^{-4}$ and $c_2 = 0.9$).

- **WW1** (a.k.a. **sufficient decrease condition** or **Armijo condition**):

$$f(x_k + \alpha_k p_k) < f(x_k) + c_1 \alpha_k \underbrace{\langle \nabla f(x_k), p_k \rangle}_{<0} =: \ell(\alpha_k).$$

  Since $\langle \nabla f(x_k), p_k \rangle < 0$, WW1 always holds for some sufficiently small α_k. When $p_k = -\nabla f(x_k)$, this is the condition we used in backtracking line search for gradient descent (Lecture 7–8).



- **WW2** (a.k.a. **curvature condition**):

$$\langle \nabla f(x_k + \alpha_k p_k), p_k \rangle \geq c_2 \underbrace{\langle \nabla f(x_k), p_k \rangle}_{<0}.$$

  Intuition: Note that the LHS equals $\phi'(\alpha_k)$ and the RHS equals $\phi'(0)$. If $\langle \nabla f(x_k + \alpha_k p_k), p_k \rangle$ is very small (very negative), then p_k is still a good descent direction as the function value $\phi(\alpha_k)$ is still decreasing, so we could keep moving along p_k by increasing α_k.

Potential downside: WW2 may hold even when $\langle \nabla f(x_k + \alpha_k p_k), p_k \rangle > 0$. If we use such an α_k, we might have moved too far. This motivates us to condition the strong Wolfe conditions.

2.1.2 Strong Wolfe conditions

Let c_1, c_2 be two numbers satisfying $0 < c_1 < c_2 < 1$.

- **Sufficient decrease condition:**

$$f(x_k + \alpha_k p_k) < f(x_k) + c_1 \alpha_k \langle \nabla f(x_k), p_k \rangle.$$

- **Curvature condition:**

$$|\langle \nabla f(x_k + \alpha_k p_k), p_k \rangle| \leq c_2 |\langle \nabla f(x_k), p_k \rangle|.$$

2.2 Existence of a good α_k

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

Proof. Let $\phi(\alpha) := f(x_k + \alpha p_k)$. By the lemma assumption, $\phi(\cdot)$ is bounded from below.

Note that the function

$$\ell(\alpha) := f(x_k) + c_1 \alpha \langle \nabla f(x_k), p_k \rangle$$

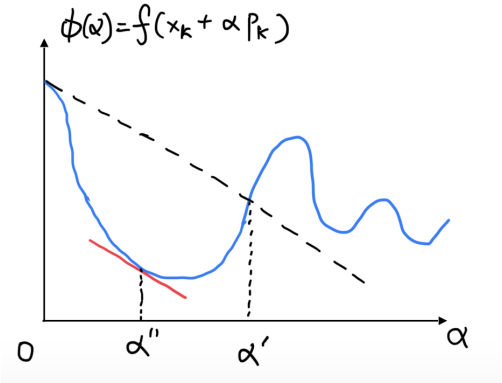
has a negative slope $c_1 \langle \nabla f(x_k), p_k \rangle < 0$. Since $c_1 < 1$ and $\phi(\cdot)$ is bounded below, there must exist $\alpha > 0$ such that $\phi(\alpha) = \ell(\alpha)$; that is, $\ell(\cdot)$ intersects $\phi(\cdot)$ at α. Let $\alpha' > 0$ be the smallest such α, with $\phi(\alpha') = \ell(\alpha')$ (see picture below). This means

$$f(x_k + \alpha' p_k) = f(x_k) + c_1 \alpha' \langle \nabla f(x_k), p_k \rangle . \tag{1}$$

Since α' is the smallest, it follows that

$$\forall \alpha < \alpha' : \quad f(x_k + \alpha p_k) < f(x_k) + c_1 \alpha \langle \nabla f(x_k), p_k \rangle .$$

So the sufficient decrease condition WW1 holds for for all $\alpha \in (0, \alpha')$.



On the other hand, by the mean-value theorem (Lecture 3, Taylor's Theorem, part 2), there exists some $\alpha'' \in (0, \alpha')$ such that

$$\begin{aligned} \phi(\alpha') &= \phi(0) + (\alpha' - 0) \cdot \phi'(\alpha'') \\ &\Updownarrow \\ f(x_k + \alpha' p_k) &= f(x_k) + \alpha' \langle \nabla f(x_k + \alpha'' p_k), p_k \rangle . \end{aligned} \tag{2}$$

See picture above for illustration. Combining (1) and (2) gives

$$\langle \nabla f(x_k + \alpha'' p_k), p_k \rangle = c_1 \underbrace{\langle \nabla f(x_k), p_k \rangle}_{<0} > c_2 \langle \nabla f(x_k), p_k \rangle \tag{3}$$

since $0 < c_1 < c_2$. Therefore, the curvature condition WW2 holds for α''. Since f is continuously differentiable, WW2 holds in a neighborhood of α'' as well.

Since the LHS and RHS of (3) are both negative, equation (3) implies

$$|\langle \nabla f(x_k + \alpha_k p_k), p_k \rangle| \leq c_2 |\langle \nabla f(x_k), p_k \rangle| ,$$

which is the curvature condition in strong Wolfe. □

2.3 Sufficient decrease + backtracking

The sufficient decrease condition WW1 alone is not sufficient to guarantee reasonable progress along the direction p_k, as α_k may be too small. However, we may use a backtracking approach to make sure that we choose a reasonably large α_k, in which case we do not need to explicitly check the curvature condition.

The following backtracking procedure generalized what we saw in Lecture 7–8. It is typically used for variants of Newton's method (but less appropriate for quasi-Newton and CG).

Algorithm 1 Backtracking Line Search

- Choose some $\bar{\alpha} > 0$ (initial value, typically $\bar{\alpha} = 1$), $\rho \in (0,1)$ (shrinkage factor), $c \in (0,1)$ (WW1 parameter)
- Set $\alpha \leftarrow \bar{\alpha}$
- repeat until WW1 is satisfied, i.e., $f(x_k + \alpha p_k) < f(x_k) + c\alpha \langle \nabla f(x_k), p_k \rangle$:

$$\text{set} \quad \alpha \leftarrow \rho\alpha$$

- return $\alpha_k = \alpha$

Besides backtracking, there are various inexact linear search algorithms for choose an α_k that satisfies the weak or strong Wolfe conditions; see Chapter 3.5 of Nocedal-Wright (optional).

2.4 Damped Newton's method

Consider the damped Newton's (DN) method

$$x_{k+1} = x_k - \alpha_k \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k), \tag{DN}$$

where α_k chosen by backtracking line search with initial value $\bar{\alpha} = 1$ and $c < 0.5$. DN converges *globally* when f is strongly convex, smooth and has Lipschitz Hessians. In particular, the convergence has two phases (HW5):

- **Damped Newton phase:** The sufficient decrease condition WW1 holds for each iteration of (DN). Following similar analysis as in gradient descent, we can show that the iterates of (DN) move towards x^* starting from an arbitrary x_0.
- **Quadratically convergent phase:** Once x_k enters a sufficiently small neighborhood of x^*, it can be shown that backtracking line search will always accept the stepsize $\bar{\alpha} = 1$. In this case, (DN) becomes the basic Newton's method and thus converges quadratically within this neighborhood.

3 Newton's method with Hessian modification

If f is nonconvex, dampening is insufficient, since $-\left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$ may not be a descent direction or even well-defined.

One solution: Modify the Hessian $\nabla^2 f(x_k)$ into some p.d. matrix $B_k \succcurlyeq \delta I$, where $\delta > 0$. Doing so ensures that $p_k = -B_k^{-1}\nabla f(x_k)$ is a descent direction: $\langle -p_k, \nabla f(x_k)\rangle = \left\langle B_k^{-1}\nabla f(x_k), \nabla f(x_k)\right\rangle \geq \lambda_{\min}\left(B_k^{-1}\right)\|\nabla f(x_k)\|_2^2$. We then use the update

$$x_{k+1} = x_k - \alpha_k B_k^{-1}\nabla f(x_k),$$

with α_k determined by a line search procedure.

Algorithm 2 Line search Newton with Hessian modification

- Input: $x_0 \in \mathbb{R}^d, \delta > 0$
- for $k = 0, 1, 2, \ldots$
  - Set $B_k = \nabla^2 f(x_k)$ if $\nabla^2 f(x_k) \succcurlyeq \delta I$; otherwise B_k is chosen so that $B_k \succcurlyeq \delta I$.
  - Compute $p_k = -B_k^{-1}\nabla f(x_k)$ (by solving the linear equation $B_k p_k = -\nabla f(x_k)$)
  - Set $x_{k+1} = x_k + \alpha_k p_k$, where α_k satisfies the Weak Wolfe Conditions.

There are several ways of choosing B_k.

- Eigenvalue modification: Suppose the Hessian has spectral decomposition

  $$\nabla^2 f(x_k) = Q\Lambda Q^\top, \qquad \Lambda = \mathrm{diag}(\vec{\lambda}).$$

  We can set

  $$B_k = Q\tilde{\Lambda}Q^\top, \qquad \text{where } \tilde{\Lambda} = \mathrm{diag}\left(\max\left\{\vec{\lambda}, \delta\mathbf{1}\right\}\right).$$

  This requires computing a full eigen decomposition of $\nabla^2 f(x_k)$, which is often too expensive for large scale problems.

- Adding a diagonal matrix: Set

  $$B_k = \nabla^2 f(x_k) + \max\left\{0, \delta - \lambda_{\min}\left(\nabla^2 f(x_k)\right)\right\} I.$$

  This requires computing/estimating the smallest eigenvalue of $\nabla^2 f(x_k)$.

- Other approaches do not use eigen decomposition, but instead compute Cholesky factorization of $\nabla^2 f(x_k)$. See Nocedal-Wright Sec 3.4 (optional).

For convergence analysis, we assume that the modified Hessian B_k satisfies the following.

Bounded modified factorization property: for all k,

$$\kappa(B_k) := \|B_k\|_2 \left\|B_k^{-1}\right\|_2 = \frac{\lambda_{\max}(B_k)}{\lambda_{\min}(B_k)} \leq C,$$

where $0 < C < \infty$. That is, the condition number of B_k is uniformly bounded for all k.

Theorem 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable, and assume that the starting point x_0 is such that the sublevel set $\mathcal{L} = \left\{x \in \mathbb{R}^d \mid f(x) \leq f(x_0)\right\}$ is compact. Then, if the bounded modified factorization property holds, we have that*

$$\lim_{k\to\infty} \nabla f(x_k) = 0.$$

Suppose the algorithm is converging to some second-order stationary point x^* with $\nabla^2 f(x^*) \succcurlyeq 2\delta I$. For all sufficiently large k, we have $\nabla^2 f(x_k) \succcurlyeq \delta I$ by continuity of $\nabla^2 f$. In this case, the algorithm will use $B_k = \nabla^2 f(x_k)$ and $\alpha_k = 1$, which becomes the basic Newton's method and enjoys local quadratic convergence.