Numerical Optimization

Lecture 9: Global convergence of line search and convergence rates of steepest descent/Newton’s methods

Sangkyun Lee

The content is from Nocedal and Wright (2006). Topics marked with ** are optional.

1 Some Background

1.1 Cauchy-Schwarz Inequality

For two vectors \( x, y \in \mathbb{R}^n \), we have

\[ |x^T y| \leq \|x\|_2 \|y\|_2 \quad \text{(Cauchy-Schwarz)} \]

** In general, for \( p, q \in [1, \infty] \) with \( 1/p + 1/q = 1 \),

\[ |x^T y| \leq \|x\|_p \|y\|_q \quad \text{(Hölder’s inequality)} \]

Note that Hölder’s inequality holds for all measurable functions \( x \) and \( y \).

1.2 Matrix Norms (Operator Norms)

Given norms on vectors \( x \in \mathbb{R}^n \) such that \( \|x\|_1 = \sum_{i=1}^{n} |x_i| \), \( \|x\|_2 = (\sum_{i=1}^{n} x_i^2)^{1/2} \), and \( \|x\|_{\infty} = \max_{i=1,...,n} |x_i| \), the **matrix norms induced by vector norms** (called operator norms) for a matrix \( A \in \mathbb{R}^{m \times n} \) is defined as

\[
\|A\| := \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{x \in \mathbb{R}^n, \|x\|=1} \|Ax\|
\]

The following can be derived from the above definition,

- \( \|A\|_1 = \max_{j=1,...,n} \sum_{i=1}^{m} |A_{ij}| \), the maximal absolute column sum.
- \( \|A\|_{\infty} = \max_{i=1,...,m} \sum_{j=1}^{n} |A_{ij}| \), the maximal absolute row sum.
- \( \|A\|_2 = \sqrt{\lambda_{\text{max}}(A^T A)} = \sigma_{\text{max}}(A) \), where \( \lambda_{\text{max}}(A^T A) \) is the largest eigenvalue of \( A^T A \) and \( \sigma_{\text{max}}(A) \) is the largest singular value of \( A \). When \( A \) is square symmetric matrix, then \( \lambda(A^T A) = \{\lambda(A)\}^2 \), so that \( \|A\|_2 = \lambda_{\text{max}}(A) \).

From the definition, it is clear that

\[ \|Ax\| \leq \|A\| \|x\|. \]

This looks like Cauchy-Schwarz inequality, but it is different since \( A \) is a matrix, not a vector.

\footnote{For a square matrix \( A \), \( \rho(A) = \max_{i=1,...,n} |\lambda_i(A)| = \|A\|_2 \) is called the spectral radius of \( A \).}
1.3 Convex Function

**Lemma 9.1.** Let $C \subseteq \mathbb{R}^n$ be a convex set and let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable over $\mathbb{R}^n$.

(a) $f$ is convex over $C$ if and only if

$$f(y) \geq f(x) + \nabla f(x)^T(y - x), \text{ for all } x, y \in C.$$

(b) $f$ is strictly convex over $C$ if any only if the above inequality is strict whenever $x \neq y$.

**Proof.** **(⇐):** Suppose that the inequality in (a) holds. Choose any $x, y \in C$ and $\alpha \in [0, 1]$ and let $z = \alpha x + (1 - \alpha) y$. Using the inequality in (a) twice, we have

$$f(x) \geq f(z) + \nabla f(z)^T(x - z)$$

$$f(y) \geq f(z) + \nabla f(z)^T(y - z).$$

Multiplying the first inequality with $\alpha$, the second by $(1 - \alpha)$, and adding them together, we obtain

$$\alpha f(x) + (1 - \alpha)f(y) \geq f(z) + \nabla f(z)^T[\alpha x + (1 - \alpha)y - z] = f(z)$$

which proves that $f$ is convex.

**(⇒):** Conversely, suppose that $f$ is convex. Let $x, y \in C$ be any vectors, and consider the function

$$g(\alpha) = \frac{f(x + \alpha(y - x)) - f(x)}{\alpha}, \quad \alpha \in (0, 1].$$

We will show that $g(\alpha)$ is monotonically increasing with $\alpha$. Consider any $\alpha_1, \alpha_2$ with $0 < \alpha_1 < \alpha_2 \leq 1$, and let

$$\bar{\alpha} = \frac{\alpha_1}{\alpha_2} \in (0, 1), \quad \bar{y} = x + \alpha_2(y - x).$$

Since $f$ is convex, we have

$$f(x + \bar{\alpha}(\bar{y} - x)) \leq \bar{\alpha} f(\bar{y}) + (1 - \bar{\alpha}) f(x)$$

That is,

$$\frac{f(x + \bar{\alpha}(\bar{y} - x)) - f(x)}{\bar{\alpha}} \leq f(\bar{y}) - f(x)$$

Plugging-in definitions of $\bar{x}$ and $\bar{\alpha}$, we get

$$\frac{f(x + \alpha_1(y - x)) - f(x)}{\alpha_1} \leq \frac{f(x + \alpha_2(y - x)) - f(x)}{\alpha_2}$$

that is,

$$g(\alpha_1) \leq g(\alpha_2).$$

This shows that $g(\alpha)$ is monotonic in $\alpha \in (0, 1]$. Then

$$(y - x)^T\nabla f(x) = \lim_{\alpha \to 0} g(\alpha) \leq g(1) = f(z) - f(x).$$

This conclude the proof for case (a). The proof for the case (b) is omitted. □
The quantity \( E(x, y) = f(y) - f(x) - \nabla f(x)^T (y - x) \) is called the excess function. Therefore \( E(x, y) \geq 0 \) for all \( x, y \in C \) for a given function \( f \) implies that \( f \) is a convex function on \( C \).

**Theorem 9.2.** Let \( f : D \subseteq \mathbb{R}^n \rightarrow \mathbb{R} \) be a twice continuously differentiable function on an open set \( D \). Then

- \( f \) is convex if and only if \( \nabla^2 f(x) \) is positive semi-definite for all \( x \in D \).
- \( f \) is strictly convex if and only if \( \nabla^2 f(x) \) is positive definite for all \( x \in D \).

**Proof.** From Taylor’s theorem, we have

\[
f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(x + \alpha(y - x))(y - x)
\]

for some \( \alpha \in [0, 1] \). When the Hessian is positive semi-definite, then

\[
E(x, y) = f(y) - f(x) - \nabla f(x^*)^T (y - x) \geq 0
\]

and therefore \( f \) is convex due to Lemma 9.1.

On the other hand, suppose that the Hessian is not p.s.d. at some point in \( D \). Since \( \nabla^2 f(x) \) is continuous, there exists \( y \in D \) such that for all \( \alpha \in [0, 1] \), \( \nabla^2 f(x + \alpha(y - x)) \) is not positive semi-definite. Then from the Taylor theorem (as above), we have

\[
E(x, y) = f(y) - f(x) - \nabla f(x)^T < 0
\]

and therefore \( f \) is not convex. The contrapositive gives that when \( f \) is convex, the Hessian is p.s.d. for all points in \( D \). \( \square \)

## 2 Global Convergence of Wolfe Line Search

Consider the following unconstrained minimization of a continuously differentiable objective function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \),

\[
\min_{x \in \mathbb{R}^n} f(x).
\]

Given the current iterate \( x_k \in \mathbb{R}^n \), \( k \geq 1 \), and a descent direction \( p_k \in \mathbb{R}^n \), we generate the next iterate \( x_{k+1} \) by

\[
x_{k+1} = x_k + \alpha_k p_k.
\]

where the step length \( \alpha_k > 0 \) can be determined by a line search strategy satisfying the Wolfe conditions (for \( 0 < c_1 < c_2 < 1 \)),

\[
\begin{align*}
f(x_k + \alpha_k p_k) & \leq f(x_k) + c_1 \alpha_k \nabla f(x_k)^T p_k \quad \text{(Sufficient Decrease)} \\
\nabla f(x_k + \alpha_k p_k)^T p_k & \geq c_2 \nabla f(x_k)^T p_k \quad \text{(Curvature)}
\end{align*}
\]

### 2.1 \( \theta_k \)

For the analysis here, the angle \( \theta_k \) between \( -\nabla f(x_k) \) and a search direction \( p_k \) plays an important role, which is defined by

\[
\cos \theta_k = \frac{-\nabla f(x_k)^T p_k}{\|\nabla f(x_k)\|_2 \|p_k\|_2}
\]
Theorem 9.3 (Global Convergence of Line Search). Consider any iteration of the form $x_{k+1} = x_k + \alpha_k p_k$, where $p_k$ is a descent direction and $\alpha_k$ satisfies the Wolfe conditions. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is bounded below in $\mathbb{R}^n$ and $f$ is continuously differentiable in an open set $\mathcal{N}$ containing the level set $\mathcal{L} := \{ x \in \mathbb{R}^n : f(x) \leq f(x_0) \}$, where $x_0$ is the starting point of the iteration. Suppose also that the gradient $\nabla f$ is Lipschitz continuous on $\mathcal{N}$, i.e. there exist a constant $L > 0$ such that

$$\| \nabla f(x) - \nabla f(y) \|_2 \leq L \| x - y \|_2, \text{ for all } x, y \in \mathcal{N}.$$ 

Then

$$\sum_{k \geq 0} \cos^2 \theta_k \| \nabla f(x_k) \|_2 < \infty \quad (\text{Zoutendijk condition}).$$

Proof. From the curvature condition and $x_{k+1} = x_k + \alpha_k p_k$, and subtracting $\nabla f(x_k)^T p_k$ from both sides, we have

$$(\nabla f(x_{k+1}) - \nabla f(x_k))^T p_k \geq (c_2 - 1) \nabla f(x_k)^T p_k$$

From the above two expressions, we obtain that

$$\alpha_k \geq \frac{(c_2 - 1) \nabla f(x_k)^T p_k}{L \| p_k \|_2^2}.$$ 

Plugging in this into the sufficient decrease condition, we get

$$f(x_{k+1}) \leq f(x_k) + c_1 \alpha_k \nabla f(x_k)^T p_k \leq f(x_k) + c_1 \frac{(c_2 - 1)(\nabla f(x_k)^T p_k)^2}{L \| p_k \|_2^2}$$

Using the definition of $\cos \theta_k$, this can be rewritten as

$$f(x_{k+1}) \leq f(x_k) - c \cos^2 \theta_k \| \nabla f(x_k) \|_2^2$$

for $c := c_1 (1 - c_2)/L > 0$. Summing up this expression for all indices up to $k$, we have

$$f(x_{k+1}) \leq f(x_0) - c \sum_{j=0}^k \cos^2 \theta_j \| \nabla f(x_j) \|_2^2.$$ 

That is, for all $k$ we have

$$\sum_{j=0}^k \cos^2 \theta_j \| \nabla f(x_j) \|_2^2 \leq \frac{1}{c} (f(x_0) - f(x_{k+1}))$$

Since $f$ is bounded below, $f(x_{k+1}) \geq -M$ for some positive scalar $M$, and therefore $f(x_0) - f(x_{k+1}) < f(x_0) + M < M'$ for some positive scalar $M'$. Hence, by taking $k \to \infty$, we get the desired result:

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \| \nabla f(x_k) \|_2^2 < \infty.$$
Note that the Zoutendijk condition implies that
\[ \lim_{k \to \infty} \cos^2 \theta_k \|\nabla f(x_k)\|_2^2 = 0. \]

When \( p_k \) is a descent direction, we have \( \cos \theta_k > 0 \) for all \( k \) (strictly speaking, we need \( \cos \theta_k \geq \delta \) \( \forall k \) for some \( \delta > 0 \), as we discuss below), and therefore the above condition implies that
\[ \lim_{k \to \infty} \|\nabla f(x_k)\|_2 = 0 \quad \text{(Global Convergence)}. \]

We call algorithms that satisfy the above condition as \textit{globally convergent}. Theorem 9.3 is the strongest global convergence result available for Wolfe line search strategy: we have the only guarantee to find a stationary point, but not necessarily a minimizer (recall however that in convex minimization a stationary point is a global minimizer).

### 2.2 Condition Number

As we’ve seen above, the condition \( \cos \theta_k > 0 \) is crucial to guarantee the global convergence of line search.

- **Steepest descent direction**: \( p_k = -\nabla f(x_k) \), so it is easy to check \( \cos \theta_k = 1 > 0 \) when \( \nabla f(x_k) \neq 0 \).

- **Newton’s direction**: \( p^N_k = -H_k^{-1} \nabla f(x_k) \) where the Hessian \( H_k = \nabla^2 f(x_k) \).

  Recall that when \( H_k \) is positive definite, we showed that \( p^N_k \) is a descent direction. Furthermore, when \( H_k \) is positive definite with a uniformly bounded condition number, that is, there exists a constant \( M > 0 \) such that
  \[ \kappa(H_k) = \|H_k\|_2 \|H_k^{-1}\|_2 \leq M, \quad \text{for all } k, \]
  where \( \kappa(H_k) \) is called the \textit{condition number} of \( H_k \). Note that if \( H_k \) is square symmetric, then \( \kappa(H_k) \) is as simple as
  \[ \kappa(H_k) = \|H_k\|_2 \|H_k^{-1}\|_2 = \frac{\lambda_n}{\lambda_1}, \]
  the ratio between the largest eigenvalue \( \lambda_n \) and the smallest eigenvalue \( \lambda_1 \) of \( H_k \). When \( \kappa(H_k) \leq M \), then we can show that
  \[ \cos \theta_k \geq \frac{1}{M} > 0 \quad \text{for all } k. \]

### 3 Rate of Convergence

#### 3.1 Convergence Rate of Steepest Descent

The behavior of steepest descent can be understood for an ideal case, in which we consider a strongly convex quadratic objective function
\[ f(x) = \frac{1}{2} x^T Q x + c^T x \]

where \( Q \in \mathbb{R}^{n \times n} \) is symmetric and positive definite. In this particular case, it is easy to perform an exact line search, finding \( \alpha > 0 \) that minimizes \( f(x_k - \alpha \nabla f(x_k)) \),
\[ f(x_k - \alpha \nabla f(x_k)) = \frac{1}{2} (x_k - \alpha \nabla f(x_k))^T Q (x_k - \alpha \nabla f(x_k)) + c^T (x_k - \alpha \nabla f(x_k)). \]
From the optimality conditions, the minimizer $\alpha_k$ is

$$\alpha_k = \frac{\nabla f(x_k)^T \nabla f(x_k)}{\nabla f(x_k)^T Q \nabla f(x_k)} > 0.$$  

(Note that if $Q$ is not positive definite, $\alpha_k$ expressed as above can have undesirable properties.) With this exact minimizer, we can write an explicit expression for the steepest descent update of minimizing strongly convex quadratic functions as

$$x_{k+1} = x_k - \frac{\nabla f(x_k)^T \nabla f(x_k)}{\nabla f(x_k)^T Q \nabla f(x_k)} \nabla f(x_k).$$

To quantify the rate of convergence, we use the weighted norm $\|x\|_Q^2 = x^T Q x$.

(Note that this definition works at best when $Q$ is positive definite.) Using the fact that an unconstrained minimizer $x^*$ of $f(x)$ satisfies $Q x^* + c = 0$, we can show that

$$\frac{1}{2} \|x - x^*\|_Q^2 = f(x) - f(x^*). \tag{9.1}$$

Using the update rule above and the fact that $g_k := \nabla f(x_k) = Q(x_k - x^*)$, we can show that

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

This can be further simplified to give the following theorem:

**Theorem 9.4 (Convergence Rate of Steepest Descent: Quadratic Case).** Using the steepest descent directions with exact line searches for minimizing the strongly convex quadratic function $f(x) = \frac{1}{2} x^T Q x + c^T x$, we have

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

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

This theorem can be shown using the Kantorovich inequality:

**Lemma 9.5 (Kantorovich Inequality).** For a positive definite symmetric $n \times n$ matrix $Q$ and for any nonzero vector $x \in \mathbb{R}^n$,

$$\frac{(x^T x)^2}{(x^T Q x)(x^T Q^{-1} x)} \geq \frac{4 \lambda_1 \lambda_n}{(\lambda_1 + \lambda_n)^2}$$

where $\lambda_1$ and $\lambda_n$ are the smallest and the largest eigenvalues of $Q$.

Few observations can be made for Theorem 9.4:

- Using (9.1), it is implied that

$$f(x_{k+1}) - f(x^*) = \left( \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \right)^2 (f(x_k) - f(x^*)).$$
• That is, the rate of convergence in terms of objective function values is linear (Lecture 2), since
\[
\frac{f(x_{k+1}) - f(x^*)}{f(x_k) - f(x^*)} = \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 \in (0, 1) \quad \forall k
\]

• The rate of convergence depends on the condition number \(\kappa(Q) = \lambda_n/\lambda_1\), since
\[
\left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 = \left(\frac{\kappa(Q) - 1}{\kappa(Q) + 1}\right)^2.
\]

• If \(\kappa(Q)\) is large (the contour of quadratic function \(f\) is elliptical), then convergence becomes slow. In fact, zigzagging can happen.

• In particular, when all eigenvalues are the same, so that \(\lambda_1 = \lambda_n\) (so that the contour of \(f\) is perfectly spherical, the steepest descent converges in a single iteration in terms of objective function values.

The convergence rate of steepest descent with exact line searches for general nonlinear objective functions is given by the following theorem:

**Theorem 9.6 (Convergence of Steepest Descent).** Let \(f : \mathbb{R}^n \to \mathbb{R}\) is twice continuously differentiable, and suppose that the iterates generated by the steepest descent with exact line searches converges to a point \(x^*\) at with \(\nabla^2 f(x^*)\) is positive definite. Let \(r\) be a scalar such that
\[
r \in \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}, 1\right)
\]
where \(\lambda_1 \leq \ldots \lambda_n\) are the eigenvalues of \(\nabla^2 f(x^*)\). Then for all sufficiently large \(k\), we have
\[
f(x_{k+1}) - f(x^*) \leq r^2(f(x_k) - f(x^*)).
\]

When inexact line searches are used (e.g. Wolfe line search), the rate of convergence does not improve in general.

### 3.2 Convergence Rate of Newton’s Method

For a twice differentiable function \(f : \mathbb{R}^n \to \mathbb{R}\), Newton directions are given by
\[
p_N^k = -(\nabla^2 f(x_k))^{-1}\nabla f(x_k).
\]
Suppose that the inverse exists. Recall that if \(\nabla^2 f(x_k)\) is not positive definite, \(p_N^k\) may not be a descent direction. There are two remedies when the Hessian is not p.d.:

• Modify the Hessian so that it becomes p.d.

• Use trust region methods.

Here we discuss the *local convergence* of Newton’s method: we know that if \(\nabla^2 f(x^*)\) is p.d., then if the Hessian \(\nabla^2 f(x)\) is continuous then \(\nabla^2 f(x)\) will be p.d. as well in a neighborhood of \(x^*\). In such a region Newton directions are well defined and descent.
**Theorem 9.7** (Local Convergence of Newton’s Method). Suppose that \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is twice differentiable and that \( \nabla^2 f(x) \) is Lipschitz continuous in a neighborhood of a solution \( x^* \) with a constant \( L > 0 \), that is,

\[
\| \nabla^2 f(x) - \nabla^2 f(y) \|_2 \leq L \| x - y \|_2 \quad \forall x, y \in N(x^*)
\]

where at \( x^* \) the second order sufficient conditions are satisfied. Consider iterations \( x_{k+1} = x_k + p_k^N \) (so that step lengths are 1). Suppose also that the starting point \( x_0 \) is sufficiently close to \( x^* \). Then,

- the sequence \( \{x_k\} \) converges quadratically to \( x^* \),
- the sequence of gradient norms \( \{\|\nabla f(x_k)\|_2\} \) converges quadratically to zero.

**Proof.**

\[
x_{k+1} - x^* = x_k + p_k^N - x^* = x_k - x^* - \nabla^2 f(x_k)^{-1} \nabla f(x_k)
\]

\[
= \nabla^2 f(x_k)^{-1}[\nabla^2 f(x_k)(x_k - x^*) - (\nabla f(x_k) - \nabla f(x^*))].
\]

Note that we have used \( \nabla f(x^*) = 0 \) from SOSC. Using the relation that

\[
\nabla f(x_k) - \nabla f(x^*) = \int_0^1 J_k \left[ \nabla f(x^* + t(x_k - x^*)) \right] dt
\]

\[
= \int_0^1 \nabla^2 f(x^* + t(x_k - x^*)) (x_k - x^*) dt,
\]

we get

\[
\|x_{k+1} - x^*\|_2
\]

\[
= \|\nabla^2 f(x_k)^{-1}[\nabla^2 f(x_k)(x_k - x^*) - (\nabla f(x_k) - \nabla f(x^*))]\|_2
\]

\[
\leq \|\nabla^2 f(x_k)^{-1}\|_2 \|\nabla^2 f(x_k)(x_k - x^*) - (\nabla f(x_k) - \nabla f(x^*))\|_2
\]

\[
= \|\nabla^2 f(x_k)^{-1}\|_2 \left\| \int_0^1 [\nabla^2 f(x_k) - \nabla^2 f(x^* + t(x_k - x^*))](x_k - x^*) dt \right\|_2
\]

\[
\leq \|\nabla^2 f(x_k)^{-1}\|_2 \int_0^1 \|\nabla^2 f(x_k) - \nabla^2 f(x^* + t(x_k - x^*))\|_2 \|x_k - x^*\|_2 dt
\]

\[
\leq \|\nabla^2 f(x_k)^{-1}\|_2 \int_0^1 \|\nabla^2 f(x_k) - \nabla^2 f(x^* + t(x_k - x^*))\|_2 \|x_k - x^*\|_2 dt
\]

\[
\leq \|\nabla^2 f(x_k)^{-1}\|_2 \|x_k - x^*\|_2 \int_0^1 L(1 - t) dt
\]

\[
= \frac{L}{2} \|\nabla^2 f(x_k)^{-1}\|_2 \|x_k - x^*\|_2^2
\]

Since \( \nabla^2 f(x^*) \) is positive definite, there exists a radius \( r > 0 \) such that \( \|\nabla^2 f(x_k)^{-1}\|_2 \leq 2\|\nabla^2 f(x^*)^{-1}\|_2 \) for all \( x_k \) with \( \|x_k - x^*\|_2 \leq r \). Together with the expression above, we obtain

\[
\|x_{k+1} - x^*\|_2 \leq L \|\nabla^2 f(x^*)^{-1}\|_2 \|x_k - x^*\|_2 \leq \tilde{L} \|x_k - x^*\|_2^2
\]

where \( \tilde{L} = L \|\nabla^2 f(x^*)^{-1}\|_2 \).
Choosing \( x_0 \) so that \( \|x_0 - x^*\|_2 \leq \min(r, 1/(2\tilde{L})) \), we can deduce from this inequality that the sequence converges to \( x^* \) (try this by yourself), where the rate of convergence is quadratic.

Finally, using the facts that \( x_{k+1} - x_k = p_k^N \) and \( \nabla f(x_k) + \nabla^2 f(x_k)p_k^N = 0 \),

\[
\|\nabla f(x_{k+1})\|_2 = \|\nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)p_k^N\|_2 \\
= \left\| \int_0^1 \nabla^2 f(x_k + tp_k^N)(x_{k+1} - x_k)dt - \nabla^2 f(x_k)p_k^N \right\|_2 \\
\leq \int_0^1 \|\nabla^2 f(x_k + tp_k^N) - \nabla^2 f(x_k)\|_2 \|p_k^N\|_2 dt \\
\leq \|p_k^N\|^2_2 \int_0^1 L t dt = \frac{L}{2}\|p_k^N\|^2_2 \\
\leq \frac{L}{2}\|\nabla^2 f(x_k)^{-1}\|^2_2 \|\nabla f(x_k)\|^2_2 \\
= \tilde{L}\|\nabla f(x_k)\|^2_2.
\]

\[\square\]

References