# Lecture 19: Basic Newton's Method

Yudong Chen

## 1 Second-Order Optimization

From now on, we will assume $\mathcal{X} = \mathbb{R}^d$ (unconstrained optimization) and $f : \mathbb{R}^d \to \mathbb{R}$ is *twice* continuously differentiable.

Second-order oracle model:

$$x \in \mathbb{R}^d \longrightarrow \boxed{\begin{array}{c}\text{2nd order}\\ \text{oracle}\end{array}} \longrightarrow f(x), \nabla f(x), \nabla^2 f(x).$$

Recall our general descent method:

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

where $\alpha_k$ is the stepsize and $p_k$ is a search direction. If $p_k$ satisfies $\langle p_k, \nabla f(x_k)\rangle < 0$, then it is called a descent direction at $x_k$.

In this and subsequent lectures, we focus on search directions of the form

$$p_k = -B_k^{-1}\nabla f(x_k),$$

where $B_k \succ 0$. Examples:

- $B_k = I$: standard gradient descent, considered before;
- $B_k = \nabla^2 f(x_k)$: Newton's method;
- $B_k =$ some approximation of $\nabla^2 f(x_k)$: quasi-Newton's methods.

## 2 Basic Newton's Method

The basic Newton's (BN) method uses $B_k = \nabla^2 f(x_k)$ with a unit stepsize $\alpha_k = 1, \forall k$. That is,

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

This (BN) update can be interpreted in two ways.

**Minimizer of second-order approximation:** When $\nabla^2 f(x_k) \succeq 0$, one can verify that

$$x_{k+1} = \operatorname*{argmin}_y \left\{ f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \frac{1}{2}\left\langle \nabla^2 f(x_k)(y - x_k), y - x_k\right\rangle \right\}.$$

We see that $x_{k+1}$ minimizes the second-order Taylor expansion of $f$ at $x_k$. (Compare this with GD.)

**Steepest descent in Hessian norm:** Using the Hessian matrix $\nabla^2 f(x)$ at $x$, one can define a weighted norm

$$\|u\|_{\nabla^2 f(x)} := \sqrt{u^\top \nabla^2 f(x) u}$$

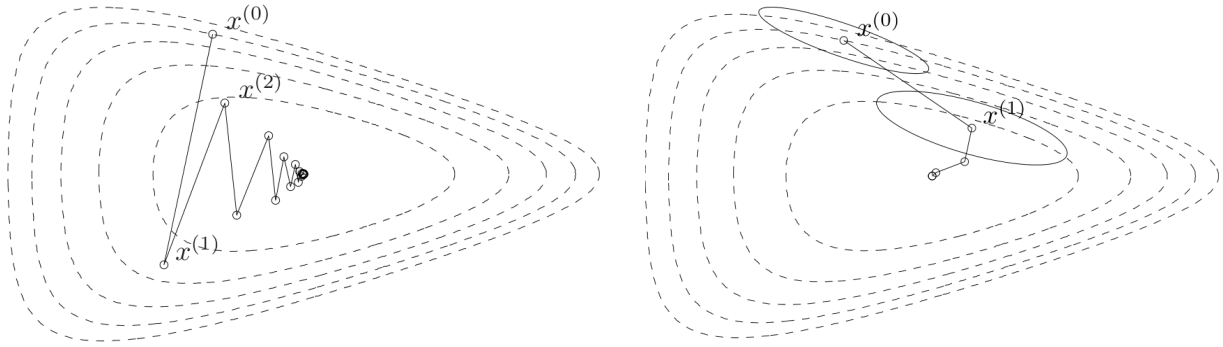
for each $u \in \mathbb{R}^d$. Define

$$p^* := \operatorname*{argmax}_{p:\|p\|_{\nabla^2 f(x_k)} \le 1} \langle -\nabla f(x_k), p \rangle .$$

It can be shown that the Newton step $p_k = -\left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$ is in the direction of $p^*$, i.e., $p_k = t p^*$ for some $t \ge 0$. That is, $p_k$ is the steepest descent direction with respect to the norm $\|\cdot\|_{\nabla^2 f(x_k)}$. Compare this with the negative gradient direction used in GD:

$$-\frac{\nabla f(x)}{\|\nabla f(x)\|_2} = \operatorname*{argmax}_{p:\|p\|_2 \le 1} \langle -\nabla f(x_k), p \rangle .$$

Below are illustrations of the steps taken by gradient descent (left) and Newton's method (right).[1] In the right plot we also show the ellipsoids $\{x : \|x - x_k\|_{\nabla^2 f(x_k)} \le 1\}$.




### 2.1 Basic assumptions

Here we assume that

- $\nabla^2 f(x_k)$ is invertible, so the iteration (BN) is well-defined;
- $\nabla^2 f(x_k) \succ 0$ is positive definite (p.d.), so $p_k = -\left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$ is a descent direction (see Lecture 6).

Later we will discuss how to handle situations where these assumptions are not satisfied.

## 3 Terminology for rates of convergence

To discuss the convergence rate of (BN) and other descent methods, we introduce some terminology.

Let $\{x_k\}$ be a sequence in $\mathbb{R}^d$ that converges to some $x^* \in \mathbb{R}^d$. We say that the convergence is

[1] The plots are taken from Convex Optimization by Boyd and Vandenberghe.

1. *Q-linear* (or simply *linear*), if there exists $r \in (0,1)$ such that

$$\|x_{k+1} - x^*\|_2 \leq r \|x_k - x^*\|_2, \qquad \forall k \text{ sufficient large.}$$

For example, the sequence $x_k = 0.7^k$ converges to 0 linearly (with $r = 0.7$). We previously showed that when $f$ is $m$-strongly convex and $L$-smooth, GD converges Q-linearly with $r \approx 1 - \frac{m}{L}$. Roughly speaking, linear convergence means that an $\epsilon$ error can be achieves in $\log \frac{1}{\epsilon}$ iterations.

2. *Q-quadratic*, if there exists a constant $M > 0$ such that

$$\|x_{k+1} - x^*\|_2 \leq M \|x_k - x^*\|_2^2, \qquad \forall k \text{ sufficient large.}$$

Note the square on the RHS. For example, the sequence $x_k = 0.7^{(2^k)}$ converges to 0 quadratically (with $M = 1$). Roughly speaking, quadratic convergence means that an $\epsilon$ error can be achieved in $\log\log \frac{1}{\epsilon}$ iterations. Put differently, the number of correct digits doubles at each iteration. Quadratic convergence is much faster than linear convergence. (A picture)

3. *Q-superlinear*, if for any constant $r > 0$, there exists $K_r < \infty$ such that

$$\|x_{k+1} - x^*\|_2 \leq r \|x_k - x^*\|_2, \qquad \forall k \geq K_r.$$

This means that $\{x_k\}$ converges faster than linear convergence with any $r$ (but not necessarily as fast as quadratic convergence).

## 4 Local quadratic convergence of Newton's method

We say that the Hessian $\nabla^2 f$ is Lipschitz-continuous with parameter $L_H < \infty$ if

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_2 \leq L_H \|x - y\|_2, \quad \forall x, y, \tag{1}$$

where on the left hand side we use the matrix operator norm (i.e., largest singular value). Recall that the condition

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \succeq mI \succ 0 \tag{2}$$

is a (2nd-order) sufficient condition for $x^*$ being a local minimizer of $f$.

The basic Newton's method converges quadratically in a neighborhood of such an $x^*$.

**Theorem 1** (Similar to Theorem 3.5 in Nocedal-Wright). *Suppose that $f$ is twice continuously differentiable, that its Hessian is $L_H$-Lipschitz-continuous, and that $x^*$ is a point satisfying the 2nd-order sufficient condition (2) for some $m > 0$. Let $\{x_k\}$ be given by (BN). If the initial point $x_0$ satisfies $\|x_0 - x^*\|_2 \leq \frac{m}{2L_H}$, then*

*(i) the sequence of iterates $\{x_k\}$ converges to $x^*$ quadratically: $\|x_{k+1} - x^*\|_2 \leq \frac{L_H}{m} \|x_k - x^*\|_2^2, \forall k$;*

*(ii) the sequence of gradient norms $\{\|\nabla f(x_k)\|_2\}$ converges to zero quadratically: $\|\nabla f(x_{k+1})\|_2 \leq \frac{2L_H}{m^2} \|\nabla f(x_k)\|_2^2, \forall k$.*

*Proof.* As our induction hypothesis, assume that $\|x_k - x^*\|_2 \leq \frac{m}{2L_H}$ .

**Proof of (i):** Recall $x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$. Then

$$\begin{aligned}
\|x_{k+1} - x^*\|_2 &= \left\| x_k - x^* - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k) \right\|_2 \\
&= \left\| \left(\nabla^2 f(x_k)\right)^{-1} \left[\nabla^2 f(x_k)\,(x_k - x^*) - \nabla f(x_k)\right] \right\|_2 \\
&\le \left\| \left(\nabla^2 f(x_k)\right)^{-1} \right\|_2 \left\| \nabla^2 f(x_k)\,(x_k - x^*) - (\nabla f(x_k) - \nabla f(x^*)) \right\|_2. \quad \text{b/c } \nabla f(x^*) = 0
\end{aligned}$$

We know from Taylor's Theorem that

$$\nabla f(x^*) - \nabla f(x_k) = \int_0^1 \nabla^2 f\left(x_k + t(x^* - x_k)\right)(x^* - x_k)\,\mathrm{d}t.$$

It follows that

$$\begin{aligned}
&\|x_{k+1} - x^*\|_2 \\
&\le \left\| \left(\nabla^2 f(x_k)\right)^{-1} \right\|_2 \left\| \int_0^1 \left[\nabla^2 f(x_k) - \nabla^2 f\left(x_k + t(x^* - x_k)\right)\right](x_k - x^*)\,\mathrm{d}t \right\|_2 \\
&\le \left\| \left(\nabla^2 f(x_k)\right)^{-1} \right\|_2 \int_0^1 \left\| \left[\nabla^2 f(x_k) - \nabla^2 f\left(x_k + t(x^* - x_k)\right)\right](x_k - x^*) \right\|_2 \mathrm{d}t \quad \text{Jensen} \\
&\le \left\| \left(\nabla^2 f(x_k)\right)^{-1} \right\|_2 \int_0^1 \underbrace{\left\| \nabla^2 f(x_k) - \nabla^2 f\left(x_k + t(x^* - x_k)\right) \right\|_2}_{\le L_H t \|x_k - x^*\|_2} \|x_k - x^*\|_2 \,\mathrm{d}t \quad \text{Cauchy-Schwarz} \\
&\le \frac{L_H}{2} \left\| \left(\nabla^2 f(x_k)\right)^{-1} \right\|_2 \|x_k - x^*\|_2^2.
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
\lambda_{\min}\left(\nabla^2 f(x_k)\right) &\ge \lambda_{\min}\left(\nabla^2 f(x^*)\right) - \left\|\nabla^2 f(x_k) - \nabla^2 f(x^*)\right\|_2 && \text{Weyl's inequality (cf. HW1 Q9.2)} \\
&\ge \lambda_{\min}\left(\nabla^2 f(x^*)\right) - L_H \|x_k - x^*\|_2 && \nabla^2 f \text{ is } L_H\text{-Lipschitz} \\
&\ge \frac{m}{2}, && \nabla^2 f(x^*) \succeq mI, \|x_k - x^*\|_2 \le \frac{m}{2L_H}
\end{aligned}$$

hence

$$\left\| \left(\nabla^2 f(x_k)\right)^{-1} \right\|_2 \le \frac{2}{m}. \tag{3}$$

Combining pieces, we obtain

$$\|x_{k+1} - x^*\|_2 \le \frac{L_H}{2} \cdot \frac{2}{m} \cdot \|x_k - x^*\|_2^2 = \frac{L_H}{m} \|x_k - x^*\|_2^2,$$

In addition, thanks to the induction hypothesis $\|x_k - x^*\|_2 \le \frac{m}{2L_H}$, we have $\|x_{k+1} - x^*\|_2 \le \frac{L_H}{m} \cdot \frac{m}{2L_H} \cdot \|x_k - x^*\|_2 = \frac{1}{2}\|x_k - x^*\|_2$, hence $\|x_k - x^*\|_2$ converges to zero. We conclude that $x_k$ converges to $x^*$ quadratically with $M = \frac{L_H}{m}$, proving (i).

Also note that $\|x_{k+1} - x^*\|_2 \le \|x_k - x^*\|_2 \le \frac{m}{2L_H}$, so the induction step is completed.

**Proof of (ii):** From $x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$, we can write

$$\nabla f(x_k) = -\nabla^2 f(x_k)\,(x_{k+1} - x_k). \tag{4}$$

Hence

$$\begin{aligned}
\|\nabla f(x_{k+1})\|_2 &= \|\nabla f(x_{k+1}) - \nabla f(x_k) + \nabla f(x_k)\|_2 \\
&= \left\| \int_0^1 \left[\nabla^2 f\left(x_k + t(x_{k+1} - x_k)\right) - \nabla^2 f(x_k)\right](x_{k+1} - x_k)\,\mathrm{d}t \right\|_2 \quad \text{Taylor and (4)} \\
&\le \int_0^1 \underbrace{\left\|\nabla^2 f\left(x_k + t(x_{k+1} - x_k)\right) - \nabla^2 f(x_k)\right\|_2}_{\le L_H t \|x_{k+1} - x_k\|_2} \|x_{k+1} - x_k\|_2 \,\mathrm{d}t \quad \text{Jensen's, Cauchy-Schwarz} \\
&\le \frac{L_H}{2} \|x_{k+1} - x_k\|_2^2 \\
&= \frac{L_H}{2} \left\| \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k) \right\|_2^2 \\
&\le \frac{L_H}{2} \cdot \underbrace{\left\| \left(\nabla^2 f(x_k)\right)^{-1} \right\|_2^2}_{\le \frac{4}{m^2} \text{ by (3)}} \cdot \|\nabla f(x_k)\|_2^2 \\
&\le \frac{2L_H}{m^2} \|\nabla f(x_k)\|_2^2 .
\end{aligned}$$

We conclude that $\|\nabla f(x_{k+1})\|_2$ converges quadratically with $M' = \frac{2L_H}{m^2}$, proving (ii). □

*Remark* 1. If $f(x) = \frac{1}{2}x^\top Ax - b^\top x$ is a convex quadratic function, then the Hessian $\nabla^2 f(x) = A$ is independent of $x$ and $\nabla^2 f$ is $L_H$-Lipschitz continuous on $\mathbb{R}^d$ with $L_H = 0$. In this case, Theorem 1 implies that (BN) converges to a global minimizer $x^*$ in one iteration. Of course, one can prove this result directly by noting that $x_1 = x_0 - A^{-1}(Ax_0 - b) = A^{-1}b = x^*$.

# 5 Additional remarks

## 5.1 Affine invariance

A nice feature of Newton's method is that it is invariant to linear or affine transformations (i.e., changes of coordinates), in the follow sense. Let $\{x_k\}$ be the iterates of (BN) applied to the function $f : \mathbb{R}^d \to \mathbb{R}$. Suppose $T \in \mathbb{R}^{d\times d}$ is a nonsingular matrix. Define a new function $g : \mathbb{R}^d \to \mathbb{R}$ by $g(y) = f(Ty)$. If we apply (BN) to minimize $g$ starting from $y_0 = T^{-1}x_0$, then

$$y_k = T^{-1}x_k, \qquad \forall k.$$

(Proof uses the chain rules $\nabla g(y) = T^\top \nabla f(Ty)$ and $\nabla^2 g(y) = T^\top \nabla^2 f(Ty)T$; left as exercise.) That is, the iterates are related by the same linear transformation. In contrast, gradient descent lacks this property and is very sensitive to changes of coordinates (which strongly affect, e.g., the condition number).

However, the convergence analysis of (BN) in Theorem 1 is *not* affine invariant: it depends very much on the choice of coordinates. If we change the coordinate system, the values of $L_H, M$ and $M'$ all change. There is an elegant way of obtaining affine invariant convergence results, which is based on the notion of self-concordant functions; see Section 6.

## 5.2 Performance (optional)

Newton's method converges very fast near $x^*$. If $x_0$ is sufficiently close to $x^*$ such that the quadratic convergence holds, usually at most six iterations suffice for achieving a very high accuracy.

The main drawback of Netwon's method is the high cost of computing and storing the $d \times d$ Hessian matrix $\nabla^2 f(x)$, especially when $d$ is large. There are several ways for reducing the computational cost, including various inexact Newton's methods and quasi-Newton's methods—we will discuss some of them later.

### 5.3 Global convergence? (optional)

Theorem 1 is a *local* convergence result: it holds when $x_0$ is sufficiently close to $x^*$. If $x_0$ is far from $x^*$, the basic Newton's method (BN) may not converge to a stationary point. Additional adjustment to (BN) is needed to ensure global convergence. We will discuss some of them in the next lecture.

## 6 Analysis of Newton's method for self-concordant functions (optional)

The "traditional" analysis of Newton's method in Theorem 1 applies to strongly convex functions with Lipschitz Hessian, and the analysis is in terms of the $\|\cdot\|_2$ norm. In this section, we present an alternative convergence analysis, discovered by Nesterov and Nemirovski, where the role of strong convexity and Lipschitz Hessian is replaced by the self-concordance property, and $\|\cdot\|_2$ is replaced by an appropriated weighted norm defined using the Hessian. This approach is simple and elegant, leading to bounds that are affine-invariant and do not depend on any unknown constants (e.g., $m$ or $L_H$).

Additional references:

- Section 9.6 in Convex Optimization by Boyd and Vandenberghe.
- Section 4.1 in Introductory Lectures on Convex Optimization by Yurii Nesterov.
- Section 5.3 in Bubeck's monograph.

### 6.1 Self-concordance functions

Assume that $f$ is three-times continuous differentiable. Note that for each $x \in \mathbb{R}^d$, the 3rd derivative $\nabla^3 f(x) \in \mathbb{R}^{d\times d\times d}$ is a 3rd order tensor. It can be equivalently viewed as a function $\nabla^3 f(x) : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ given by $\nabla^3 f(x)[a,b,c] = \sum_{i,j,k} \left(\nabla^3 f(x)\right)_{ijk} a_i b_j c_k$.[2] The previous Lipschitz Hessian assumption (1) is equivalent to

$$\nabla^3 f(x)[h,h,h] \le L_H \|h\|_2^3, \qquad \forall x \in \mathbb{R}^d, h \in \mathbb{R}^d.$$

The idea of self-concordance is to replace $\|h\|_2$ on the RHS by the Hessian norm $\|h\|_x \equiv \|h\|_{\nabla^2 f(x)} := \sqrt{h^\top \nabla^2 f(x) h}$, which we introduced earlier.

**Definition 1** (Self-concordance)**.** A three-times continuously differentiable function is (standard) *self-concordant* if

$$\nabla^3 f(x)[h,h,h] \le 2\|h\|_x^3, \qquad \forall x \in \mathbb{R}^d, h \in \mathbb{R}^d.$$

In $d = 1$ dimension, the above definition is equivalent to $|f'''(x)| \le 2\left(f''(x)\right)^{3/2}, \forall x$. For general $d$, $f$ is self-concordant if it is self-concordant along every 1-D line, i.e., the 1-D function $\tilde{f}_v(t) := f(x+tv)$ is self-concordant for every $v \in \mathbb{R}^d$.

Examples of self-concordant functions:

[2] This is analogous to viewing the Hessian matrix $\nabla^2 f(x) \in \mathbb{R}^{d\times d}$ as a bilinear function $\nabla^2 f(x) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ given by $\nabla^2 f(x)[a,b] = \sum_{i,j} \left(\nabla^2 f(x)\right)_{ij} a_i b_j$.

- Linear and quadratic functions, for which $\nabla^3 f(x) = 0$.
- Negative log $f(x) = -\log x$, and the log barrier function $f(x) = -\sum_{i=1}^n \log(b_i - a_i^\top x)$.
- Log-determinant: $f(X) = -\log\det X$, defined for p.d. matrices $X \succ 0$.

In a similar spirit, instead of measuring progress by $\|\nabla f(x)\|_2$ (done in Theorem 1), we use the *dual* Hessian norm.

**Definition 2** (Newton decrement)**.** Let $f$ be a strictly convex self-concordant function. The *Newton decrement* at $x$ is defined as

$$\lambda(x) := \|\nabla f(x)\|_x^* = \sqrt{\nabla f(x)^\top \left(\nabla^2 f(x)\right)^{-1} \nabla f(x)},$$

where $\|\cdot\|_x^*$ denotes the dual norm of the Hessian norm $\|\cdot\|_x$.

Let $x^*$ be a minimizer of $f$. When $f$ is self-concordant, the Newton decrement controls the distance to $x^*$ and the optimality gap. In particular, for all $x$ with $\lambda(x) \le 0.68$, we have

$$\begin{aligned} \|x - x^*\|_x &\le \frac{\lambda(x)}{1-\lambda(x)}, \\ f(x) - f(x^*) &\le \lambda(x)^2. \end{aligned}$$

These are analogs of the following inequalities for $m$-strongly convex functions $f$:

$$\begin{aligned} \|x - x^*\|_2 &\le \frac{1}{m}\|\nabla f(x)\|_2, \\ f(x) - f(x^*) &\le \frac{1}{2m}\|\nabla f(x)\|_2^2. \end{aligned}$$

## 6.2 Convergence analysis

Consider the basic Newton's method (BN), i.e., $x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$, applied to self-concordant functions. We have the following beautiful quadratic convergence result.

**Theorem 2.** *Let $f$ be a self-concordant strictly convex function. If $x_0$ satisfies $\lambda(x_0) \le \frac{1}{4}$, then*

$$\lambda(x_{k+1}) \le 2\lambda(x_k)^2, \quad \forall k \ge 0.$$

This theorem is an analog of Theorem 1. Notably, it does not depend on any unknown parameters of the function $f$.

To prove Theorem 2, we derive some basic properties of self-concordance. First consider a one-dimensional strictly convex function $\tilde{f} : \mathbb{R} \to \mathbb{R}$, for which self-concordance means $\left|\tilde{f}'''(t)\right| \le 2\left(\tilde{f}''(t)\right)^{3/2}, \forall t \in \mathbb{R}$. This is in turn equivalent to

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\left(\tilde{f}''(t)^{-1/2}\right)\right| \le 1, \quad \forall t \in \mathbb{R}.$$

Integrating the above inequality from 0 to $t$ gives

$$-t \le \tilde{f}''(t)^{-1/2} - \tilde{f}''(0)^{-1/2} \le t, \quad \forall t \ge 0,$$

from which we can obtain

$$\frac{\tilde{f}''(0)}{\left(1+t\tilde{f}''(0)^{1/2}\right)^2} \le \tilde{f}''(t) \le \frac{\tilde{f}''(0)}{\left(1-t\tilde{f}''(0)^{1/2}\right)^2}, \quad \forall 0 \le t < \tilde{f}''(0)^{-1/2}. \tag{5}$$

Now consider a $d$-dimensional self-concordant strictly convex function $f$. With some work, the above inequality (5) can be generalized to this $f$:

$$(1-t\|v\|_x)^2 \nabla f^2(x) \preceq \nabla^2 f(x+tv) \preceq \frac{1}{(1-t\|v\|_x)^2} \nabla f^2(x), \qquad \forall 0 \le t < \frac{1}{\|v\|_x}, \forall v \in \mathbb{R}^d. \tag{6}$$

*Proof of Theorem 2.* Fix an arbitrary $x \in \mathbb{R}^d$, and set $v = -\left(\nabla^2 f(x)\right)^{-1} \nabla f(x)$. Then $x^+ = x + v$ is the output after a Newton step from $x$, and $\|v\|_x = \lambda(x)$. When $\lambda(x) \le \frac{1}{4}$, we can apply the lower bound in (6) to obtain

$$\nabla^2 f(x^+) \succeq (1-\lambda(x))^2 \nabla^2 f(x) \implies \left[\nabla^2 f(x^+)\right]^{-1} \preceq \frac{1}{(1-\lambda(x))^2} \left[\nabla^2 f(x)\right]^{-1}.$$

It follows that

$$\begin{aligned} \lambda(x^+) &= \sqrt{\nabla f(x^+)^\top \left[\nabla^2 f(x^+)\right]^{-1} \nabla f(x^+)} \\ &\le \sqrt{\nabla f(x^+)^\top \frac{\left[\nabla^2 f(x)\right]^{-1}}{(1-\lambda(x))^2} \nabla f(x^+)} = \frac{1}{1-\lambda(x)} \left\|\nabla f(x^+)\right\|_x^*. \end{aligned} \tag{7}$$

On the other hand, by definition of $x^+ = x + v$ and Taylor's theorem, we have the expression

$$\begin{aligned} \nabla f(x^+) &= \nabla f(x^+) - \nabla f(x) - \nabla^2 f(x)(x^+ - x) \\ &= \Big\langle \underbrace{\int_0^1 \left[\nabla^2 f(x+tv) - \nabla^2 f(x)\right] \mathrm{d}t}_{G}, v \Big\rangle = Gv \end{aligned}$$

Hence it holds that

$$\begin{aligned} &\left\|\nabla f(x^+)\right\|_x^{*2} \\ &= v^\top G \left[\nabla^2 f(x)\right]^{-1} Gv \\ &= \left(\left[\nabla^2 f(x)\right]^{1/2} v\right)^\top \cdot \underbrace{\left[\nabla^2 f(x)\right]^{-1/2} G \left[\nabla^2 f(x)\right]^{-1/2}}_{H} \cdot \underbrace{\left[\nabla^2 f(x)\right]^{-1/2} G \left[\nabla^2 f(x)\right]^{-1/2}}_{H} \cdot \left[\nabla^2 f(x)\right]^{1/2} v \\ &\le \|H\|_2^2 \cdot \|v\|_x^2 = \|H\|_2^2 \cdot \lambda(x)^2. \end{aligned}$$

By integrating the bound (6), we have

$$\left(-\lambda(x) + \frac{1}{3}\lambda(x)^2\right) \nabla^2 f(x) \preceq G \preceq \frac{\lambda(x)}{1-\lambda(x)} \nabla^2 f(x),$$

hence $\|H\|_2 \le \max\left\{\frac{\lambda(x)}{1-\lambda(x)}, \lambda(x) - \frac{1}{3}\lambda(x)^2\right\} = \frac{\lambda(x)}{1-\lambda(x)}$. It follows that

$$\left\|\nabla f(x^+)\right\|_x^{*2} \le \frac{\lambda(x)^2}{(1-\lambda(x))^2} \lambda(x)^2. \tag{8}$$

Combining (7) and (8) gives

$$\lambda(x^+) \leq \frac{1}{1-\lambda(x)} \cdot \frac{\lambda(x)^2}{1-\lambda(x)} \leq 4\lambda(x)^2 \qquad \text{when } \lambda(x) \leq \frac{1}{4}.$$

Applying this bound to $x = x_k$ and using induction on $k$, we prove Theorem 2. □