Part 6. Newton’s Method

Math 126 Winter 18

Date of current version: February 28, 2018

Abstract This note studies Newton’s methods. Many parts of this note are based on the chapters [1, Chapter 5] [2, Chapters 9-11] [3, Chapters 1,4].

Please email me if you find any typos or errors.

1 Newton’s Method (see [1, Chapter 5] [2, Chapter 9] [3, Chapter 1])

Consider the following unconstrained minimization problem:

$$\min_{x} f(x),$$

where $f$ is twice continuously differentiable.

1.1 Newton’s Method

The update of the Newton’s method is given by

$$x_{k+1} = \arg \min_{x} \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} (x - x_k)^T \nabla^2 f(x_k) (x - x_k) \right\},$$

which reduces to the following when we further assume that $\nabla^2 f(x_k)$ is positive definite:

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k).$$

Remark 1.1 Newton’s direction $d = -[\nabla^2 f(x)]^{-1} \nabla f(x) \neq 0$ is not necessarily a descent direction, i.e.,

$$\nabla f(x)^T d = -\nabla f(x)^T \nabla^2 f(x) \nabla f(x) \neq 0.$$

If $\nabla^2 f(x)$ is positive definite, then Newton’s direction is a descent direction.

At the very least, Newton’s method requires that $\nabla^2 f(x)$ is positive definite for every $x \in \mathbb{R}^d$, which in particular implies that there exists a unique optimal solution $x^*$. However, this is not enough to guarantee convergence, as the following example illustrates.

Example 1.1 Consider $f(x) = \sqrt{1 + x^2}$ defined over the real line. The minimizer of $f$ over $\mathbb{R}$ is $x^* = 0$. The first and second derivatives of $f$ are

$$f'(x) = \frac{x}{\sqrt{1 + x^2}}, \quad f''(x) = \frac{1}{(1 + x^2)^{3/2}}.$$
The update of Newton’s method has the form
\[ x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} = x_k - x_k(1 + x_k^2) = -x_k^3. \]

We therefore see that for \( |x_0| \geq 1 \) the method diverges and that for \( |x_0| < 1 \) the method converges very rapidly to the solution \( x_\ast = 0 \).

**Theorem 1.1 (quadratic local convergence of Newton’s method)** Let \( f \) be a twice continuously differentiable function defined over \( \mathbb{R}^d \). Assume that
- there exists \( \mu > 0 \) for which \( \nabla^2 f(x) \geq \mu I \) for any \( x \in \mathbb{R}^d \),
- there exists \( M > 0 \) for which \( \|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq M\|x - y\|_2 \) for any \( x, y \in \mathbb{R}^d \).

Let \( \{x_k\}_{k \geq 0} \) be the sequence generated by Newton’s method, and let \( x_\ast \) be unique minimizer of \( f \) over \( \mathbb{R}^d \). Then
\[
\|x_{k+1} - x_\ast\|_2 \leq \frac{M}{2\mu} \|x_k - x_\ast\|_2^2. \tag{1.5}
\]

In addition, if \( \|x_0 - x_\ast\|_2 \leq \frac{\mu}{M} \), then
\[
\|x_k - x_\ast\|_2 \leq \frac{2\mu}{M} \left( \frac{1}{2} \right)^{\frac{k}{2}}. \tag{1.6}
\]

**Proof** By the fundamental theorem of calculus, we have
\[
x_{k+1} - x_\ast = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k) - x_\ast
= x_k - x_\ast + [\nabla^2 f(x_k)]^{-1} (\nabla f(x_k) - \nabla f(x_\ast))
= x_k - x_\ast + [\nabla^2 f(x_k)]^{-1} \int_0^1 \nabla^2 f(x_k + t(x_\ast - x_k))(x_\ast - x_k) dt
= [\nabla^2 f(x_k)]^{-1} \int_0^1 \nabla^2 f(x_k + t(x_\ast - x_k)) - \nabla^2 f(x_k)(x_\ast - x_k) dt.
\]

Then,
\[
\|x_{k+1} - x_\ast\|_2 \leq \|\nabla^2 f(x_k)^{-1}\|_2 \sup_{t \in [0,1]} \left\| \int_0^1 \nabla^2 f(x_k + t(x_\ast - x_k)) - \nabla^2 f(x_k)(x_\ast - x_k) dt \right\|_2
\leq \|\nabla^2 f(x_k)^{-1}\|_2 \int_0^1 \left\| \nabla^2 f(x_k + t(x_\ast - x_k)) - \nabla^2 f(x_k)(x_\ast - x_k) \right\|_2 dt
\leq \|\nabla^2 f(x_k)^{-1}\|_2 \int_0^1 \|\nabla^2 f(x_k + t(x_\ast - x_k)) - \nabla^2 f(x_k)\|_2 \|x_\ast - x_k\|_2 dt
\leq \frac{1}{\mu} \int_0^1 M t \|x_k - x_\ast\|_2^2 dt
= \frac{M}{2\mu} \|x_k - x_\ast\|_2^2.
\]

We next prove the second inequality (1.6) by induction. Note that for \( k = 0 \), we assumed that
\[
\|x_0 - x_\ast\|_2 \leq \frac{\mu}{M} = \frac{2\mu}{M} \left( \frac{1}{2} \right)^0.
\]

Assume that (1.6) holds for an integer \( k \), that is \( \|x_k - x_\ast\|_2 \leq \frac{2\mu}{M} \left( \frac{1}{2} \right)^k \); we will show it holds for \( k + 1 \). By (1.5) we have
\[
\|x_{k+1} - x_\ast\|_2 \leq \frac{M}{2\mu} \|x_k - x_\ast\|_2^2 \leq \frac{M}{2\mu} \left( \frac{2\mu}{M} \left( \frac{1}{2} \right)^k \right)^2 = \frac{2\mu}{M} \left( \frac{1}{2} \right)^{k+1}.
\]

\( \Box \)

**Remark 1.2** An iterative method is called locally convergent if the generated sequence converges to an optimal point \( x_\ast \) given that the initial point \( x_0 \) is close enough to \( x_\ast \).
1.2 Damped Newton’s Method

Newton’s method does not guarantee descent of the function values even when the Hessian is positive definite, similar to a gradient method with step size $s_k = 1$, i.e. $x_{k+1} = x_k - \nabla f(x_k)$. This can be fixed by introducing a step size chosen by a certain line search, leading to the following damped Newton’s method.

**Algorithm 1 Damped Newton’s Method**

1: Input: $x_0 \in \mathbb{R}^d$.
2: for $k \geq 0$ do
3: \hspace{1em} Compute the Newton direction $d_k$, which is the solution to the linear system $\nabla^2 f(x_k)d_k = -\nabla f(x_k)$.
4: \hspace{1em} Choose a step size $s_k > 0$ using a backtracking line search.
5: \hspace{1em} $x_k = x_k + s_k d_k$.
6: \hspace{1em} If a stopping criteria is satisfied, then stop.

**Remark 1.3** Backtracking line search: starting from an initial $s > 0$, repeat $s \leftarrow \beta s$ until the following sufficient decrease condition is satisfied:

$$f(x + sd) < f(x) + \alpha s \nabla f(x)^\top d$$

with parameters $\alpha \in (0, 1)$ and $\beta \in (0, 1)$.

There exists constants $\eta \in \left(0, \frac{\mu^2}{M^2}\right)$, and $\gamma > 0$ such that the following holds; specifically, $0 < \eta \leq \min \left\{\frac{\mu^2}{M^2}, 3(1 - 2\alpha)\frac{\mu^2}{M^2}\right\}$ and $\gamma = \alpha \beta \eta^2 \frac{L}{M}$ for a constant $L$ satisfying $\nabla^2 f(x) \preceq LI$ for any $x \in \mathbb{R}^d$.

- (damped Newton phase) If $\|\nabla f(x_k)\|_2 \geq \eta$, then
  $$f(x_{k+1}) - f(x_k) \leq -\gamma.$$  

- (quadratically convergent phase) If $\|\nabla f(x_k)\|_2 < \eta$, then the backtracking line search selects $s_k = 1$ and
  $$\|\nabla f(x_{k+1})\|_2 \leq \frac{M^2}{2\mu} \|\nabla f(x_k)\|_2^2.$$  

**Remark 1.4** Unlike the gradient method that is affected by changes of coordinates, Newton’s method is independent of affine changes of coordinates. Suppose $T \in \mathbb{R}^{d \times d}$ is nonsingular, and define $f(y) = f(Ty)$. If we use Newton’s method (with the same backtracking parameters) to minimize $f$, starting from $y_0 = T^{-1}x_0$, then we have $Ty_k = x_k$ for all $k$.

1.3 Self-concordance

Previous convergence analysis involves the constants $\mu$, $M$, and $L$, which are difficult to estimate in practice. In addition, even though Newton’s method is affinely invariant, the previous convergence analysis is not affinely invariant.

Consider the following assumption, called self-concordance, that does not depend on any unknown constants (such as $\mu$, $M$, and $L$) and affine changes of coordinates.

**Definition 1.1** A convex function $f : \mathbb{R} \rightarrow \mathbb{R}$ is self-concordant if

$$|f'''(x)| \leq \kappa f''(x)^{\frac{3}{2}}$$

for all $x \in \text{dom } f$, where $\kappa$ is some positive constant. $\kappa = 2$ is a standard choice.

A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is self-concordant if it is self-concordant along every line in its domain, i.e., if the function $f(t) = f(x + tv)$ is a self-concordant function of $t$ for all $x \in \text{dom } f$ and for all $v$.

**Remark 1.5** The self-concordant functions include many of the logarithmic barrier functions that play an important role in interior point methods for solving convex optimization problems.
Example 1.2 Self-concordant functions.

– Negative logarithm $f(x) = -\log x$: Using $f''(x) = \frac{1}{x^2}$, $f'''(x) = -\frac{2}{x^3}$, we have
  \[
  \frac{|f'''(x)|}{2f''(x)^{\frac{3}{2}}} = 1.
  \]

– Negative entropy plus negative logarithm $f(x) = x \log x - \log x$

Self-concordance is preserved under

– scaling by a factor $a \geq 1$,
– addition,
– composition with affine function,
– composition with logarithm: Let $g : \mathbb{R} \to \mathbb{R}$ be a convex function with $\text{dom } g = \mathbb{R}_{++}$ and
  \[
  |g'''(x)| \leq 3 \frac{g''(x)}{x}
  \]
  for all $x$. Then
  \[
  f(x) = -\log(-g(x)) - \log x
  \]
  is self-concordant on $\{x : x > 0, g(x) < 0\}$

Example 1.3 Self-concordant functions.

– Log barrier for linear inequalities $f(x) = -\sum_{i=1}^{m} \log(b_i - a_i^T x)$ on $\{x : a_i^T x < b_i, i = 1, \ldots, m\}$
– Log-determinant $f(X) = -\log \det X$ on $\mathbb{S}^d_{++}$
– $f(x, y) = -\log(y^2 - x^T x)$ on $\{(x, y) : ||x||_2 \leq y\}$.

For strictly convex self-concordant function, we obtain bounds in terms of the Newton decrement

\[
\lambda(x) = \sqrt{\nabla f(x)^T [\nabla^2 f(x)]^{-1} \nabla f(x)}.
\]

There exists constants $\eta \in (0, 1/4]$ and $\gamma > 0$ (that depend only on the backtracking line search parameters $\alpha$ and $\beta$) such that the following holds.

– (damped Newton phase) If $\lambda(x_k) > \eta$, then
  \[
  f(x_{k+1}) - f(x_k) \leq -\gamma.
  \]
– (quadratically convergent phase) If $\lambda(x_k) \leq \eta$, then the backtracking line search selects $s_k = 1$ and
  \[
  \lambda(x_{k+1}) \leq 2\lambda(x_k)^2
  \]
2 Newton’s Method with Equality Constraints (see [2, Chapter 10])

Consider the equality constrained convex minimization problem:

$$\min_{x \in \mathbb{R}^d} f(x)$$

subject to $Ax = b,$

where $f : \mathbb{R}^d \to \mathbb{R}$ is convex and twice continuously differentiable, and $A \in \mathbb{R}^{p \times d}$ with $\text{rank}(A) = p < d.$

To derive the Newton direction $\Delta x_{nt}$ at the feasible point $\hat{x}$, we solve the following second-order Taylor approximation near $\hat{x}$:

$$\min_{v \in \mathbb{R}^d} \left\{ \hat{f}(\hat{x} + v) = f(\hat{x}) + \langle \nabla f(\hat{x}), v \rangle + \frac{1}{2}v^T \nabla^2 f(\hat{x})v \right\}$$

subject to $A(\hat{x} + v) = b.$

Recall that a point $v_* \in \mathbb{R}^d$ is optimal for (2.2) if and only there is a $\mu_* \in \mathbb{R}^p$ such that

$$\nabla_v \hat{f}(\hat{x} + v_*) + A^T \mu_* = \nabla f(\hat{x}) + \nabla^2 f(\hat{x}) v_* + A^T \mu_* = 0, \quad A(\hat{x} + v_*) = b,$$

which is a linearized approximation of the optimality conditions of (2.1) near $\hat{x}$:

$$\nabla f(x_*) + A^T \mu_* = 0, \quad Ax_* = b.$$  (2.4)

Using $Ax = b,$ the Newton direction $\Delta x_{nt} = v_*$ can be simply characterized by

$$\begin{pmatrix} \nabla^2 \hat{f}(\hat{x}) & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} v_* \\ \mu_* \end{pmatrix} = \begin{pmatrix} -\nabla f(\hat{x}) \\ 0 \end{pmatrix}.$$  (2.5)

The Newton direction is defined only at points for which the KKT matrix is nonsingular.

On the other hand, one can parameterize the feasible set $\{x : Ax = b\}$ as

$$\{x : Ax = b\} = \{Fz + \tilde{x} : z \in \mathbb{R}^{d-p}\}$$  (2.6)

with a feasible point $\tilde{x}$ that satisfies $A\tilde{x} = b$ and a matrix $F \in \mathbb{R}^{d \times (d-p)}$ whose range is the nullspace of $A.$ Then, one can eliminate the equality constraints of the problem (2.1) as

$$\min_{z \in \mathbb{R}^{d-p}} \left\{ \hat{f}(z) = f(Fz + \tilde{x}) \right\}. $$  (2.7)

The corresponding Newton’s direction is

$$\Delta z_{nt} = -[\nabla^2 \hat{f}(z)]^{-1} \nabla \hat{f}(z) = -[F^T \nabla^2 f(x)F]^{-1} F^T \nabla f(x),$$

where $x = Fz + \tilde{x}.$ One can show that (proof omitted)

$$\Delta x_{nt} = F\Delta z_{nt}.$$  (2.9)

Then, starting from $x_0 = Fz_0 + \tilde{x},$ the iterates $\{x_k\}_{k \geq 0}$ and $\{z_k\}_{k \geq 0}$ of Newton’s method for (2.1) and (2.7) respectively satisfy

$$x_k = Fz_k + \tilde{x},$$

and thus the convergence analysis of Newton’s method for unconstrained problem directly applies to Newton’s method solving the equality constrained problems.

What if we don’t have a feasible point $\tilde{x}?$ A Newton’s direction $\Delta x$ can be found by solving

$$\begin{pmatrix} \nabla^2 f(x) & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \mu \end{pmatrix} = -\begin{pmatrix} \nabla f(x) \\ Ax - b \end{pmatrix}.$$  (2.11)
3 Newton’s Method with Inequality Constraints (see [2, Chapter 11] [3, Chapter 4])

Consider the equality and inequality constrained convex minimization problem:

\[
\min_{x \in \mathbb{R}^d} f(x) \quad \text{(3.1)}
\]

subject to \( g_i(x) \leq 0, \ i = 1, \ldots, m, \)

\[ Ax = b, \]

where \( f, g_i : \mathbb{R}^d \to \mathbb{R} \) are convex and twice continuously differentiable, and \( A \in \mathbb{R}^{p \times d} \) with \( \text{rank}(A) = p < d \). We also assume that there exists \( \hat{x} \in \text{dom} f \) that satisfies \( A\hat{x} = b \) and \( g_i(\hat{x}) < 0 \) for \( i = 1, \ldots, m \).

We approximate the problem into an equality constrained problem to which Newton’s method can be applied. We first use an indication function to make the inequality constraints implicit as:

\[
\min_{x \in \mathbb{R}^d} f(x) + \sum_{i=1}^{m} I_{\mathbb{R}_+}(g_i(x)) \quad \text{(3.2)}
\]

subject to \( Ax = b \).

The objective function is yet non-differentiable, so Newton’s method cannot be applied.

The basic idea of the barrier method, a particular interior-point method, is to approximate the indicator function \( I_{\mathbb{R}_+} \) by

\[
I_{\mathbb{R}_+}(u) = I_{\mathbb{R}_+}(-u) \approx \hat{I}_-(u) = -\frac{1}{t} \log(-u) \quad \text{(3.3)}
\]

with \( t > 0 \), where \( \text{dom} \hat{I}_- = -\mathbb{R}_+ \). The approximation improves as \( t \to \infty \). Substituting \( \hat{I}_- \) for \( I_{\mathbb{R}_+} \) leads to

\[
\min_{x \in \mathbb{R}^d} \left\{ f(x) - \sum_{i=1}^{m} \frac{1}{t} \log(-g_i(x)) \right\} \quad \text{(3.4)}
\]

subject to \( Ax = b \),

which can be solved by Newton’s method.

The function

\[
\psi(x) = -\sum_{i=1}^{m} \log(-g_i(x)) \quad \text{(3.5)}
\]

is called the logarithmic barrier (or interior penalty), which is convex and twice continuously differentiable.

For \( t > 0 \), define \( x_*(t) \) as the solution of

\[
\min_{x \in \mathbb{R}^d} \{ tf(x) + \psi(x) \} \quad \text{(3.6)}
\]

subject to \( Ax = b \),

which can be solved via Newton’s method. The central path associated with the problem (3.6) is defined as \( \{ x_*(t) : t > 0 \} \), a set of central points \( x_*(t) \).

Central points \( x_*(t) \) for \( t > 0 \) are characterized by the following necessary and sufficient conditions: \( x_*(t) \) is strictly feasible for (3.1), i.e., satisfies

\[
g_i(x_*(t)) < 0, \ i = 1, \ldots, m, \quad Ax_*(t) = b, \quad \text{(3.7)}
\]

and there exists \( \mu \) such that

\[
0 = tf(x_*(t)) + \nabla \psi(x_*(t)) + A^\top \mu \quad \text{(3.8)}
\]

\[
= tf(x_*(t)) + \sum_{i=1}^{m} \left( -\frac{1}{g_i(x_*(t))} \right) \nabla g_i(x_*(t)) + A^\top \mu.
\]
We can then state that every central point $x_\ast (t)$ yields a dual feasible point $(\lambda_\ast (t), \mu_\ast (t))$:

$$\lambda_{\ast, i}(t) = -\frac{1}{l g_i(x_\ast (t))}, \quad i = 1, \ldots, m, \quad \mu_\ast (t) = \frac{\mu}{l},$$

(3.9)

and hence a lower bound on the primal optimal value $p_\ast$. Specifically, from (3.8), a central point $x_\ast (t)$ minimizes the Lagrangian of (3.1) for $(\lambda_\ast (t), \mu_\ast (t))$:

$$L(x, \lambda_\ast (t), \mu_\ast (t)) = f(x) + \sum_{i=1}^{m} \lambda_{\ast, i}(t) g_i(x) + \mu_\ast (t)^\top (A x - b).$$

(3.10)

Then the dual function is

$$q(\lambda_\ast (t), \mu_\ast (t)) = \inf_x L(x, \lambda_\ast (t), \mu_\ast (t)) = f(x_\ast (t)) - \frac{m}{l},$$

(3.11)

which implies

$$f(x_\ast (t)) - p_\ast \leq f(x_\ast (t)) - g(\lambda_\ast (t), \mu_\ast (t)) = \frac{m}{l}.$$  

(3.12)

This confirms our intuition that $x_\ast (t)$ converges to an optimal point as $t \to \infty$.

**Algorithm 2** Barrier Method (Path-following Method)

1: **Input:** strictly feasible $x_0 \in \mathbb{R}^d$, $t_0 > 0$, $\rho > 1$.
2: for $k \geq 0$ do
3: Compute $x_\ast (t_k)$ by minimizing $tf(x) + \psi(x)$ subject to $A x = b$ starting at $x_k$ (via Newton’s method).
4: $x_{k+1} = x_\ast (t_k)$
5: if a stopping criteria is satisfied, then stop.
6: $t_{k+1} = \rho t_k$.

**References**