Newton's Method for Optimization

Yinyu Ye

Department of Management Science and Engineering

Stanford University

Stanford, CA 94305, U.S.A.

http://www.stanford.edu/~yyye

LY: Chapter 10

Newton's method for unconstrained optimization

All unconstrained local minimizers of a differentiable function f are KKT or stationary points: they make the gradient ∇f vanish. Thus, finding a solution of the KKT condition

$$\nabla f(\mathbf{x}) = \mathbf{0}$$

is a matter of solving a system of (possibly nonlinear) equations.

For functions of a single real variable, the KKT condition is

$$f'(x) = 0.$$

When f is twice continuously differentiable, Newton's method can be a very effective way to solve such equations and hence to locate a stationary point of f.

Note that Newton's method is a procedure for solving equations, and it would be a second-order method for optimization.

Newton's method for solving the equation $g(x) = 0$

The univariate case. Given a starting point x^0, Newton's method for solving the equation $g(x) = 0$ is to generate the sequence of iterates

$$x^{k+1} = x^k - \frac{g(x^k)}{g'(x^k)}.$$

The iteration is well defined provided that $g'(x^k) \neq 0$ at each step.

Newton's method for solving a system of equations $\mathbf{g}(\mathbf{x}) = \mathbf{0}$

When we have a mapping

$$\mathbf{g}(\mathbf{x}) = \begin{bmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_n(\mathbf{x}) \end{bmatrix},$$

we define the Jacobian of $\mathbf{g}$ as

$$\nabla \mathbf{g}(\mathbf{x}) = \left[\frac{\partial g_i(\mathbf{x})}{\partial x_j}\right].$$

The rows of $\nabla \mathbf{g}(\mathbf{x})$ are the gradient vectors

$$\nabla g_1(\mathbf{x}), \ldots, \nabla g_n(\mathbf{x}).$$

For the system $\mathbf{g}(\mathbf{x}) = \mathbf{0}$, i.e.,

$$g_i(\mathbf{x}) = 0, \quad i = 1, \ldots, n$$

the iteration is given by

$$\mathbf{x}^{k+1} = \mathbf{x}^k - (\nabla \mathbf{g}(\mathbf{x}^k))^{-1} \mathbf{g}(\mathbf{x}^k).$$

This formula follows from the use of a Taylor series approximation to g at the point x^k, namely

$$\mathbf{g}(\mathbf{x}) \approx \mathbf{g}(\mathbf{x}^k) + \nabla \mathbf{g}(\mathbf{x}^k)(\mathbf{x} - \mathbf{x}^k).$$

When we set the right-hand side of this equation to zero, we can solve it for $\mathbf{x}$, provided that the Jacobian matrix is nonsingular.

Newton's method for minimizing $f(\mathbf{x})$

For optimization, we target at $\nabla f(\mathbf{x}) = \mathbf{0}$ so that the iteration is given by

$$\mathbf{x}^{k+1} = \mathbf{x}^k - (\nabla^2 f(\mathbf{x}^k))^{-1} \nabla f(\mathbf{x}^k).$$

Improvement Measures: the quality of iterates can be measures by $\|\mathbf{x} - \mathbf{x}^*\|$ ($\mathbf{x}^*$ is an KKT point), $\|\nabla f(\mathbf{x})\|$, or often by

$$\|\nabla f(\mathbf{x})\|_{(\nabla^2 f(\mathbf{x}))^{-1}} = \sqrt{\nabla f(\mathbf{x})^T (\nabla^2 f(\mathbf{x}))^{-1} \nabla f(\mathbf{x})},$$

where $\nabla^2 f(\mathbf{x})$ is positive definite.

Example: for a given constant a to minimize

$$a \cdot x - \log(x), \quad x > 0$$

. Also, see Problem 5 of HW3.

Local Convergence Theorem

Theorem 1 *Let $f(\mathbf{x})$ be twice continuously differentiable and satisfy the (second-order) β-Lipschitz condition, that is, for any two points $\mathbf{x}$ and $\mathbf{y}$*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) + \nabla^2 f(\mathbf{y})(\mathbf{x} - \mathbf{y})\| \leq \beta \|\mathbf{x} - \mathbf{y}\|^2$$

for a positive real number β. Also let $\mathbf{x}^*$ be a local minimizer of f at which $\nabla^2(\mathbf{x}^*)$ is positive definite. Then, provided that $\|\mathbf{x}^0 - \mathbf{x}^*\|$ is sufficiently small, the sequence generated by Newton's method converges quadratically to $\mathbf{x}^*$.

$$
\begin{aligned}
\|\mathbf{x}^{k+1}-\mathbf{x}^*\| &= \|\mathbf{x}^k-\mathbf{x}^*-\nabla^2 f(\mathbf{x}^k)^{-1}\nabla f(\mathbf{x}^k)\| \\
&= \|\nabla^2 f(\mathbf{x}^k)^{-1}\left(\nabla f(\mathbf{x}^k)-\nabla^2 f(\mathbf{x}^k)(\mathbf{x}^k-\mathbf{x}^*)\right)\| \\
&= \|\nabla^2 f(\mathbf{x}^k)^{-1}\left(\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^*)-\nabla^2 f(\mathbf{x}^k)(\mathbf{x}^k-\mathbf{x}^*)\right)\| \\
&\le \|\nabla^2 f(\mathbf{x}^k)^{-1}\|\|\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^*)-\nabla^2 f(\mathbf{x}^k)(\mathbf{x}^k-\mathbf{x}^*)\| \\
&\le \|\nabla^2 f(\mathbf{x}^k)^{-1}\|\beta\|\mathbf{x}^k-\mathbf{x}^*\|^2 \\
&\le c\|\mathbf{x}^k-\mathbf{x}^*\|^2,
\end{aligned} \tag{1}
$$

for some constant c. Thus, when $c\|\mathbf{x}^0-\mathbf{x}^*\|<1$, the quadratic convergence takes place.

Is Newton's direction descent ?

For an arbitrary twice-continuously differentiable function f, there is no reason to expect Newton's method produce a sequence of iterates that converge to a local minimizer of f. In fact, as we have seen, convergence to anything is not guaranteed without additional hypotheses. Then, how do we modify the Newton method?

If our search direction at a point, say $\bar{\mathbf{x}}$ is

$$\mathbf{d} = -(\nabla^2 f(\bar{\mathbf{x}}))^{-1} \nabla f(\bar{\mathbf{x}})^T,$$

then it is a descent direction for the objective function only if

$$\nabla f(\bar{\mathbf{x}})\mathbf{d} = -\nabla f(\bar{\mathbf{x}})(\nabla^2 f(\bar{\mathbf{x}}))^{-1} \nabla f(\bar{x})^T < 0$$

which will hold if $\nabla^2 f(\bar{\mathbf{x}})$ is a positive definite matrix, but it is only a sufficient condition, however.

The Quasi-Newton Method

In general:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha^k S^k \nabla f(\bar{\mathbf{x}})^T,$$

for a symmetric matrix S^k with a step-size scalar α^k.

SDM: $S^k = I$, α^k is decided by line search.

Newton: $S^k = (\nabla^2 f(\mathbf{x}^k))^{-1}$, $\alpha^k = 1$ or by line search (see Problem 5 of HW3).

Hibrid: $S^k = (\nabla^2 f(\mathbf{x}^k) + \lambda I)^{-1}$, $\alpha^k = 1$ or by line search.

Various methods were developed such that $S^0 = I$, and then S^k is gradually becoming $(\nabla^2 f(\mathbf{x}^k))^{-1}$; see Chapter 10.

Some computational issues

As an optimization technique, Newton's method, in its pure form, requires knowledge of (or the capacity to compute) the first and second derivatives of the objective function. In an environment where individual function evaluations are expensive, this could be a drawback.

In a large-scale problem, the computation of the search direction, $\mathbf{d}$, could also turn out to be a time-consuming task. One may solve $S^k\mathbf{d} = -\nabla f(\mathbf{x}^k)$ using matrix factorization methods. Thus, the symmetric LDL^T factorization could be used to compute $\mathbf{d}$:

$$LDL^T = S^k.$$

Then, compute the search direction:

$$LDL^T\mathbf{d} = -\nabla f(\mathbf{x}^k)^T.$$

Typically, we reorder the variables such that L is sparse.

An application case of Newton's method

Consider the optimization problem

$$\begin{array}{ll} \min & -\sum_j \ln x_j \\ \text{s.t.} & A\mathbf{x} - \mathbf{b} = \mathbf{0} \in R^m, \\ & \mathbf{x} \geq \mathbf{0}. \end{array}$$

Note this is a (strict) convex optimization problem. Suppose the feasible region has an interior and it is bounded, then the (unique) minimizer is called the analytic center of the feasible region, and it, together with multipliers $\mathbf{y}, \mathbf{s}$, satisfy the following optimality conditions:

$$\begin{array}{rcl} x_j s_j & = & 1,\ j=1,...,n, \\ A\mathbf{x} & = & \mathbf{b}, \\ A^T\mathbf{y}+\mathbf{s} & = & \mathbf{0}, \\ (\mathbf{x},\mathbf{s}) & \geq & \mathbf{0}. \end{array}$$

Since the inequality $(\mathbf{x},\mathbf{s}) \geq \mathbf{0}$ would not be active, this is a system $2n+m$ equations of $2n+m$ variables: (using $X = \text{Diag}(\mathbf{x})$)

$$\begin{array}{rcl} X\mathbf{s}-\mathbf{e} & = & 0,\ j=1,...,n, \\ A\mathbf{x}-\mathbf{b} & = & \mathbf{0}, \\ A^T\mathbf{y}+\mathbf{s} & = & \mathbf{0}. \end{array} \qquad (2)$$

Thus, Newton's method would be applicable...

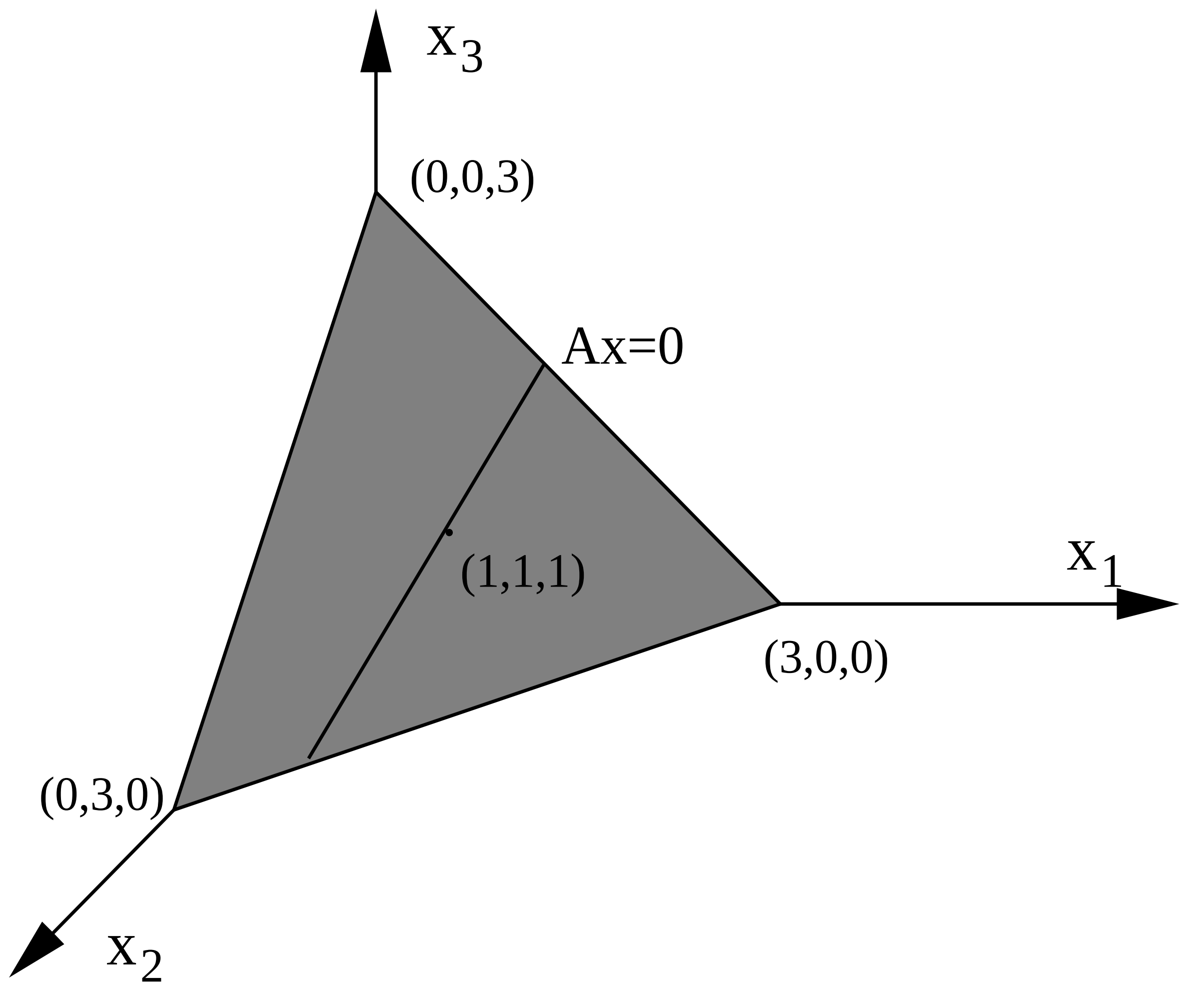

Figure 1: Illustration of the primal polytope and its analytic center.

Newton Direction

Let $(\mathbf{x} > \mathbf{0}, \mathbf{y}, \mathbf{s} > \mathbf{0})$ be an initial point. Then, the Newton direction would be solution of the following linear equations:

$$\begin{pmatrix} S & \mathbf{0} & X \\ A & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & A^T & I \end{pmatrix} \begin{pmatrix} \mathbf{d}_x \\ \mathbf{d}_y \\ \mathbf{d}_s \end{pmatrix} = \begin{pmatrix} \mathbf{e} - X\mathbf{s} \\ \mathbf{b} - A\mathbf{x} \\ -A^T\mathbf{y} - \mathbf{s} \end{pmatrix}.$$

Note that after one Newton iteration, the error residuals of the second and third equations vanishes. Thus, we may assume that the initial point satisfies

$$A\mathbf{x} = \mathbf{b}, \; A^T\mathbf{y} + \mathbf{s} = \mathbf{0}$$

and they remain satisfied through out the process.

Newton Direction Simplification

$$\begin{aligned} S\mathbf{d}_x + X\mathbf{d}_s &= \mathbf{e} - X\mathbf{s}, \\ A\mathbf{d}_x &= \mathbf{0}, \\ A^T\mathbf{d}_y + \mathbf{d}_s &= \mathbf{0}. \end{aligned} \tag{3}$$

Multiplying AS^{-1} to the top equation and noting $A\mathbf{d}_x = \mathbf{0}$, we have

$$AXS^{-1}\mathbf{d}_s = AS^{-1}(\mathbf{e} - X\mathbf{s}),$$

which together with the third equation give

$$\mathbf{d}_y = -(AXS^{-1}A^T)^{-1}AS^{-1}(\mathbf{e} - X\mathbf{s}),$$

$$\mathbf{d}_s = -A^T\mathbf{d}_y, \quad \text{and} \quad \mathbf{d}_x = S^{-1}(\mathbf{e} - X\mathbf{s} - X\mathbf{d}_s).$$

The new Newton iterate would be

$$\mathbf{x}^+ = \mathbf{x} + \mathbf{d}_x, \quad \mathbf{y}^+ = \mathbf{y} + \mathbf{d}_y, \quad \mathbf{s}^+ = \mathbf{s} + \mathbf{d}_s.$$

Approximate Centers

The error residual of the first equation would be:

$$\eta(\mathbf{x}, \mathbf{s}) := \|X\mathbf{s} - \mathbf{e}\|. \tag{4}$$

We now prove the following theorem

Theorem 2 *If the starting point of the Newton procedure satisfies* $\eta(\mathbf{x}, \mathbf{s}) < 2/3$*, then*

$$\mathbf{x}^+ > \mathbf{0}, \quad A\mathbf{x}^+ = \mathbf{b}, \quad \mathbf{s}^+ = \mathbf{c}^T - A^T\mathbf{y}^+ > \mathbf{0}$$

and

$$\eta(\mathbf{x}^+, \mathbf{s}^+) \le \frac{\sqrt{2}\eta(\mathbf{x}, \mathbf{s})^2}{4(1 - \eta(\mathbf{x}, \mathbf{s}))} .$$

Proof:

To prove the result we first see that

$$\|X^+\mathbf{s}^+ - \mathbf{e}\| = \|D_x\mathbf{d}_s\|, \quad D_x = \text{Diag}(\mathbf{d}_x).$$

Multiplying the both sides of the first equation of (3) by $(XS)^{-1/2}$, we see

$$D\mathbf{d}_x + D^{-1}\mathbf{d}_s = \mathbf{r} := (XS)^{-1/2}(\mathbf{e} - X\mathbf{s}),$$

where $D = S^{1/2}X^{-1/2}$. Let $\mathbf{p} = D\mathbf{d}_x$ and $\mathbf{q} = D^{-1}\mathbf{d}_s$. Note that $\mathbf{p}^T\mathbf{q} = \mathbf{d}_x^T\mathbf{d}_s = 0$ and $\mathbf{p} + \mathbf{q} = \mathbf{r}$. Then,

$$\begin{aligned} \|D_x\mathbf{d}_s\|^2 &= \|P\mathbf{q}\|^2 \\ &= \sum_{j=1}^{n}(p_j q_j)^2 \end{aligned}$$

$$
\begin{aligned}
&\leq \left(\sum_{p_j q_j > 0}^{n} p_j q_j\right)^2 + \left(\sum_{p_j q_j < 0} p_j q_j\right)^2 \\
&= 2\left(\sum_{p_j q_j > 0}^{n} p_j q_j\right)^2 \\
&\leq 2\left(\sum_{p_j q_j > 0}^{n} (p_j + q_j)^2/4\right)^2 \\
&\leq 2\left(\|\mathbf{r}\|^2/4\right)^2 .
\end{aligned}
$$

Furthermore,

$$
\|\mathbf{r}\|^2 \leq \|(XS)^{-1/2}\|^2 \|\mathbf{e} - X\mathbf{s}\|^2 \leq \frac{\eta^2(\mathbf{x}, \mathbf{s})}{1 - \eta(\mathbf{x}, \mathbf{s})} ,
$$

which gives the desired result. We leave the proof of $\mathbf{x}^+, \mathbf{s}^+ > \mathbf{0}$ as an Exercise.