

Second Order Optimization Algorithms I

Yinyu Ye

Department of Management Science and Engineering

Stanford University

Stanford, CA 94305, U.S.A.

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

Chapters 8.6-7, 9.1-9.5, 10.1-4

The 1.5-Order Algorithm: Dimension-Reduced Second-Order Method

Let $\mathbf{d}^k = \mathbf{x}^k - \mathbf{x}^{k-1}$ and $\mathbf{g}^k = \nabla f(\mathbf{x}^k)$ be two (conjugate) descent directions, and Hessian $H^k = \nabla^2 f(\mathbf{x}^k)$. Then the step-sizes can be chosen from the two-dimensional Newton method:

$$\begin{pmatrix} (\mathbf{g}^k)^T H^k \mathbf{g}^k & -(\mathbf{d}^k)^T H^k \mathbf{g}^k \\ -(\mathbf{d}^k)^T H^k \mathbf{g}^k & (\mathbf{d}^k)^T H^k \mathbf{d}^k \end{pmatrix} \begin{pmatrix} \alpha^g \\ \alpha^m \end{pmatrix} = \begin{pmatrix} \|\mathbf{g}^k\|^2 \\ -(\mathbf{g}^k)^T \mathbf{d}^k \end{pmatrix}.$$

Then, let

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

If the Hessian $\nabla^2 f(\mathbf{x}^k)$ is not available, one can approximate

$$H^k \mathbf{g}^k \sim \nabla(\mathbf{x}^k + \mathbf{g}^k) - \mathbf{g}^k \quad \text{and} \quad H^k \mathbf{d}^k \sim -(\nabla f(\mathbf{x}^k - \mathbf{d}^k) - \nabla f(\mathbf{x}^k)) = -(\mathbf{g}^{k-1} - \mathbf{g}^k).$$

For convex quadratic minimization, the method becomes the conjugate-gradient method or parallel-tangent method.

The 1.5-Order Algorithm: Quasi-Newton Method I

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

for a symmetric matrix S^k with a step-size α^k . When S^k is a nonnegative diagonal matrix, then it is the **scaled** steepest descent method we described earlier. In general, when S^k is positive definite, direction $-S^k \nabla f(\mathbf{x}^k)$ is a **descent direction** (why?).

For convex quadratic minimization, the linear convergence rate then becomes $\left(\frac{\lambda_{\max}(S^k Q) - \lambda_{\min}(S^k Q)}{\lambda_{\max}(S^k Q) + \lambda_{\min}(S^k Q)} \right)^2$ where λ_{\max} and λ_{\min} represent the largest and smallest eigenvalues of a matrix.

Thus, S^k can be viewed as a **Preconditioner**—typically an approximation of the Hessian matrix inverse, and can be learned from a regression model: let $\mathbf{p}^k = \mathbf{x}^{k+1} - \mathbf{x}^k = \alpha^k \mathbf{d}^k$

$$\mathbf{q}^k := \mathbf{g}(\mathbf{x}^{k+1}) - \mathbf{g}(\mathbf{x}^k) = Q(\mathbf{x}^{k+1} - \mathbf{x}^k) = Q\mathbf{p}^k, \quad k = 0, 1, \dots$$

We actually learn Q^{-1} from $Q^{-1}\mathbf{q}^k = \mathbf{p}^k$, $k = 0, 1, \dots$. The process start with H^k , $k = 0, 1, \dots$, where the rank of H^k is k , that is, we each step learn a rank-one update: given H^{k-1} , \mathbf{q}^k , \mathbf{p}^k we solve $(h_0 \cdot H^{k-1} + \mathbf{h}^k(\mathbf{h}^k)^T)\mathbf{q}^k = \mathbf{p}^k$ for vector \mathbf{h}^k . Then after n iterations, we build up $H^n = Q^{-1}$.

The 1.5-Order Algorithm: Quasi-Newton Method II

One can simply let

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha^k \left(\frac{n-k}{n} I + \frac{k}{n} H^k \right) \mathbf{g}(\mathbf{x}^k),$$

which is similar to the Conjugate Gradient method.

A popular method, BFGS, is given as follows (there are multiple typos in the text): start from \mathbf{x}^0 and set $S^0 = I$, let

$$\mathbf{d}^k = -S^k \mathbf{g}(\mathbf{x}^k) = -S^k \nabla f(\mathbf{x}^k),$$

and iterate

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha^k \mathbf{d}^k.$$

Then update

$$S^{k+1} = S^k + \left(1 + \frac{(\mathbf{q}^k)^T S^k \mathbf{q}^k}{(\mathbf{p}^k)^T \mathbf{q}^k} \right) \frac{\mathbf{p}^k (\mathbf{p}^k)^T}{(\mathbf{p}^k)^T \mathbf{q}^k} - \frac{\mathbf{p}^k (\mathbf{q}^k)^T S^k + S^k \mathbf{q}^k (\mathbf{p}^k)^T}{(\mathbf{p}^k)^T \mathbf{q}^k}.$$

The 1.5-Order Algorithm: The Ellipsoid Method

Ellipsoids are just sets of the form

$$E = \{\mathbf{y} \in \mathbf{R}^m : (\mathbf{y} - \bar{\mathbf{y}})^T B^{-1}(\mathbf{y} - \bar{\mathbf{y}}) \leq 1\}$$

where $\bar{\mathbf{y}} \in \mathbf{R}^m$ is a given point (called the **center**) and B is a symmetric **positive definite** matrix of dimension m . We can use the notation $\text{ell}(\bar{\mathbf{y}}, B)$ to specify the ellipsoid E defined above. Note that

$$\text{vol}(E) = (\det B)^{1/2} \text{vol}(S(\mathbf{0}, 1)).$$

where $S(\mathbf{0}, 1)$ is the unit sphere in \mathbf{R}^m .

A Half-Ellipsoid

By a **Half-Ellipsoid** of E , we mean the set

$$\frac{1}{2}E_a := \{\mathbf{y} \in E : \mathbf{a}^T \mathbf{y} \leq \mathbf{a}^T \bar{\mathbf{y}}\}$$

for a given non-zero vector $\mathbf{a} \in \mathbf{R}^m$ where $\bar{\mathbf{y}}$ is the **center** of E – the intersection of the ellipsoid and a plane cutting through the center.

We are interested in finding a new ellipsoid containing $\frac{1}{2}E_a$ with the least volume.

- How small could it be?
- How easy could it be constructed?

The New Containing Ellipsoid

The new ellipsoid $E^+ = \text{ell}(\bar{\mathbf{y}}^+, B^+)$ can be constructed as follows. Define

$$\tau := \frac{1}{m+1}, \quad \delta := \frac{m^2}{m^2-1}, \quad \sigma := 2\tau.$$

And let

$$\bar{\mathbf{y}}^+ := \bar{\mathbf{y}} - \frac{\tau}{(\mathbf{a}^T B \mathbf{a})^{1/2}} B \mathbf{a},$$

$$B^+ := \delta \left(B - \sigma \frac{B \mathbf{a} \mathbf{a}^T B}{\mathbf{a}^T B \mathbf{a}} \right).$$

Theorem 1 Ellipsoid $E^+ = \text{ell}(\bar{\mathbf{y}}^+, B^+)$ defined as above is the ellipsoid of *least volume* containing $\frac{1}{2}E_a$. Moreover,

$$\frac{\text{vol}(E^+)}{\text{vol}(E)} \leq \exp \left(-\frac{1}{2(m+1)} \right)$$

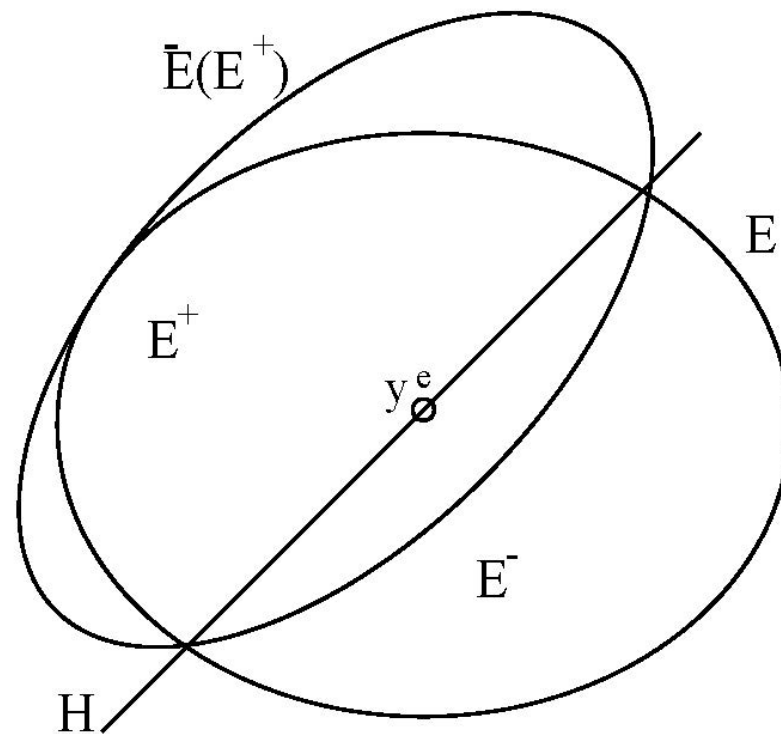


Figure 1: The least volume ellipsoid containing a half ellipsoid

The Ellipsoid Method for Minimizing a Convex Function

Consider $\min_{\mathbf{x}} f(\mathbf{x})$:

- **Initialization:** Set the initial ellipsoid (ball) as $B^0 = \frac{1}{R^2} I$ centered at an initial solution \mathbf{x}^0 where R is sufficiently large such it contains an optimal solution.
- For $k = 0, 1, \dots$ do
 - If not terminated:
 - Compute the (sub)gradient vector $\nabla f(\mathbf{x}^k)$,
 - Let the cutting-plane be $\{\mathbf{x} : \nabla f(\mathbf{x}^k)^T \mathbf{x} \leq f(\mathbf{x}^k)^T \mathbf{x}^k\}$ and form the half ellipsoid; and update \mathbf{x}^k and B^k as described earlier.

Newton's Method: The Second Order Method

For **multi-variables**, Newton's method for minimizing $f(\mathbf{x})$ is to minimize the second-order Taylor expansion function at point \mathbf{x}^k :

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

We now introduce the second-order β -**Lipschitz** condition: for any point \mathbf{x} and direction vector \mathbf{d}

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

which implies

$$f(\mathbf{x} + \mathbf{d}) - f(\mathbf{x}) \leq \nabla f(\mathbf{x})^T \mathbf{d} + \frac{1}{2} \mathbf{d}^T \nabla^2 f(\mathbf{x}) \mathbf{d} + \frac{\beta}{3} \|\mathbf{d}\|^3.$$

In the following, for notation simplicity, we use $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})$ and $\nabla \mathbf{g}(\mathbf{x}) = \nabla^2 f(\mathbf{x})$. Thus,

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

Indeed, Newton's method was initially developed for solving a system of nonlinear equations in the form $\mathbf{g}(\mathbf{x}) = \mathbf{0}$.

Local Convergence Theorem of Newton's Method

Theorem 2 Let $f(\mathbf{x})$ be β -Lipschitz and the smallest absolute eigenvalue of its Hessian uniformly bounded below by $\lambda_{min} > 0$. Then, provided that $\|\mathbf{x}^0 - \mathbf{x}^*\|$ is sufficiently small, the sequence generated by Newton's method converges quadratically to \mathbf{x}^* that is a KKT solution with $\mathbf{g}(\mathbf{x}^*) = \mathbf{0}$.

$$\begin{aligned}
 \|\mathbf{x}^{k+1} - \mathbf{x}^*\| &= \|\mathbf{x}^k - \mathbf{x}^* - \nabla \mathbf{g}(\mathbf{x}^k)^{-1} \mathbf{g}(\mathbf{x}^k)\| \\
 &= \|\nabla \mathbf{g}(\mathbf{x}^k)^{-1} (\mathbf{g}(\mathbf{x}^k) - \nabla \mathbf{g}(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*))\| \\
 &= \|\nabla \mathbf{g}(\mathbf{x}^k)^{-1} (\mathbf{g}(\mathbf{x}^k) - \mathbf{g}(\mathbf{x}^*) - \nabla \mathbf{g}(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*))\| \\
 &\leq \|\nabla \mathbf{g}(\mathbf{x}^k)^{-1}\| \|\mathbf{g}(\mathbf{x}^k) - \mathbf{g}(\mathbf{x}^*) - \nabla \mathbf{g}(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*)\| \\
 &\leq \|\nabla \mathbf{g}(\mathbf{x}^k)^{-1}\| \beta \|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{\beta}{\lambda_{min}} \|\mathbf{x}^k - \mathbf{x}^*\|^2.
 \end{aligned} \tag{1}$$

Thus, when $\frac{\beta}{\lambda_{min}} \|\mathbf{x}^0 - \mathbf{x}^*\| < 1$, the quadratic convergence takes place:

$$\frac{\beta}{\lambda_{min}} \|\mathbf{x}^{k+1} - \mathbf{x}^*\| \leq \left(\frac{\beta}{\lambda_{min}} \|\mathbf{x}^k - \mathbf{x}^*\| \right)^2.$$

Such a starting solution \mathbf{x}^0 is called an approximate root of $\mathbf{g}(\mathbf{x})$.

An application case of Newton's method

Consider the optimization problem

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

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 **y**, **s**, satisfy the following optimality conditions:

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

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 $X = \text{Diag}(\mathbf{x})$)

$$\begin{aligned} X\mathbf{s} - \mathbf{e} &= \mathbf{0}, \\ A\mathbf{x} - \mathbf{b} &= \mathbf{0}, \\ A^T\mathbf{y} + \mathbf{s} &= \mathbf{0}. \end{aligned} \tag{2}$$

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

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

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

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

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

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:

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

We now prove the following theorem

Theorem 3 *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

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

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,

$$\begin{aligned} \|D_x \mathbf{d}_s\|^2 &= \|P\mathbf{q}\|^2 \\ &= \sum_{j=1}^n (p_j q_j)^2 \\ &\leq \left(\sum_{p_j q_j > 0}^n p_j q_j \right)^2 + \left(\sum_{p_j q_j < 0} p_j q_j \right)^2 \end{aligned}$$

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

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.

Spherical Constrained Nonconvex Quadratic Minimization I

$$\min \frac{1}{2} \mathbf{x}^T Q \mathbf{x} + \mathbf{c}^T \mathbf{x}, \quad \text{s.t.} \quad \|\mathbf{x}\|^2 = (\leq) 1.$$

where $Q \in S^n$ is any symmetric data matrix. If $\mathbf{c} = \mathbf{0}$ this problem becomes finding the least eigenvalue of Q .

The necessary and sufficient condition (can be proved using the SDP Rank Theorem) for \mathbf{x} being a global minimizer of the problem is

$$(Q + \lambda I)\mathbf{x} = -\mathbf{c}, \quad (Q + \lambda I) \succeq \mathbf{0}, \quad \|\mathbf{x}\|_2^2 = 1,$$

which implies $\lambda \geq -\lambda_{\min}(Q) > 0$ where $\lambda_{\min}(Q)$ is the least eigenvalue of Q . If the optimal $\lambda^* = -\lambda_{\min}(Q)$, then \mathbf{c} must be orthogonal to the $\lambda_{\min}(Q)$ -eigenvector, and it can be checked using the power algorithm.

The minimal objective value:

$$\frac{1}{2} \mathbf{x}^T Q \mathbf{x} + \mathbf{c}^T \mathbf{x} = -\frac{1}{2} \mathbf{x}^T (Q + \lambda I) \mathbf{x} - \frac{1}{2} \lambda \|\mathbf{x}\|^2 \leq -\frac{\lambda}{2}, \quad (5)$$

Sphere Constrained Nonconvex Quadratic Minimization II

WLOG, Let us assume that the least eigenvalue is 0. Then we must have $\lambda \geq 0$. If the optimal $\lambda^* = 0$, then \mathbf{c} must be a 0-eigenvector of Q , and it can be checked using the power algorithm to find it. Therefore, we assume that the optimal $\lambda > 0$.

Furthermore, there is an upper bound on λ :

$$\lambda \leq \lambda \|\mathbf{x}\|^2 \leq \mathbf{x}^T (Q + \lambda I) \mathbf{x} = -\mathbf{c}^T \mathbf{x} \leq \|\mathbf{c}\| \|\mathbf{x}\| = \|\mathbf{c}\|.$$

Now let $\mathbf{x}(\lambda) = -(Q + \lambda I)^{-1} \mathbf{c}$, the problem becomes finding the root of $\|\mathbf{x}(\lambda)\|^2 = 1$.

Lemma 1 *The analytic function $\|\mathbf{x}(\lambda)\|^2$ is convex monotonically decreasing with $\alpha = 12$ in Corollary 1 of Lecture-Slide Note 9.*

Theorem 4 *The 1-spherical constrained quadratic minimization can be computed in $O(\log \log(\|\mathbf{c}\|/\epsilon))$ iterations where each iteration solve a symmetric (positive definite) system of linear equations of n variables.*

What about 2-spherical constrained quadratic minimization, that is, quadratic minimization with 2 ellipsoidal constraints: Remains Open.

Spherical Trust-Region Method for Minimizing Lipschitz $f(\mathbf{x})$

Recall the second-order β -Lipschitz condition: for any two points \mathbf{x} and \mathbf{d}

$$\|\mathbf{g}(\mathbf{x} + \mathbf{d}) - \mathbf{g}(\mathbf{x}) - \nabla \mathbf{g}(\mathbf{x})\mathbf{d}\| \leq \beta \|\mathbf{d}\|^2,$$

where $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})$ and $\nabla \mathbf{g}(\mathbf{x}) = \nabla^2 f(\mathbf{x})$. It implies

$$f(\mathbf{x} + \mathbf{d}) - f(\mathbf{x}) \leq \nabla f(\mathbf{x})^T \mathbf{d} + \frac{1}{2} \mathbf{d}^T \nabla^2 f(\mathbf{x}) \mathbf{d} + \frac{\beta}{3} \|\mathbf{d}\|^3.$$

$$\begin{aligned} & f(\mathbf{x} + \mathbf{d}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T \mathbf{d} - \frac{1}{2} \mathbf{d}^T \nabla^2 f(\mathbf{x}) \mathbf{d} \\ &= \int_0^1 \mathbf{d}^T (\nabla f(\mathbf{x} + t\mathbf{d}) - \nabla f(\mathbf{x})) dt - \frac{1}{2} \mathbf{d}^T \nabla^2 f(\mathbf{x}) \mathbf{d} \\ &= \int_0^1 \mathbf{d}^T (\nabla f(\mathbf{x} + t\mathbf{d}) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})(t\mathbf{d})) dt \\ &\leq \int_0^1 \|\mathbf{d}\| \|\nabla f(\mathbf{x} + t\mathbf{d}) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})(t\mathbf{d})\| dt \\ &\leq \int_0^1 \|\mathbf{d}\| \beta \|t\mathbf{d}\|^2 dt \text{ (by 2nd-order -Lipschitz condition)} \\ &= \beta \|\mathbf{d}\|^3 \int_0^1 t^2 dt = \frac{\beta}{3} \|\mathbf{d}\|^3. \end{aligned}$$

The second-order method, at the k th iterate, would let $\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{d}^k$ where

$$\begin{aligned} \mathbf{d}^k = & \arg \min_{\mathbf{d}} \quad (\mathbf{c}^k)^T \mathbf{d} + \frac{1}{2} \mathbf{d}^T Q^k \mathbf{d} + \frac{\beta}{3} \alpha^3 \\ \text{s.t.} \quad & \|\mathbf{d}\| \leq \alpha, \end{aligned}$$

with $\mathbf{c}^k = \nabla f(\mathbf{x}^k)$ and $Q^k = \nabla^2 f(\mathbf{x}^k)$. One typically fixed α to a “trusted” radius α^k so that it becomes a sphere-constrained problem (the inequality is normally active if the Hessian is non PSD):

$$(Q^k + \lambda^k I) \mathbf{d}^k = -\mathbf{c}^k, \quad (Q^k + \lambda^k I) \succeq \mathbf{0}, \quad \|\mathbf{d}^k\|_2^2 = (\alpha^k)^2.$$

For fixed α^k , the method is generally called **trust-region** method.

The Trust-Region can be ellipsoidal such as $\|S\mathbf{d}\| \leq \alpha$ where S is a PD diagonal scaling matrix.

Convergence Speed of the Spherical Trust-Region Method

Is there a trusted radius such that the method converging? A simple choice would fix $\alpha^k = \sqrt{\epsilon}/\beta$. Then from reduction (5)

$$f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k) \leq -\frac{\lambda^k}{2} \|\mathbf{d}^k\|^2 + \frac{\beta}{3} (\alpha^k)^3 = -\frac{\lambda^k (\alpha^k)^2}{2} + \frac{\beta}{3} (\alpha^k)^3 = -\frac{\lambda^k \epsilon}{2\beta^2} + \frac{\epsilon^{3/2}}{3\beta^2}.$$

Also

$$\begin{aligned} \|\mathbf{g}(\mathbf{x}^{k+1})\| &= \|\mathbf{g}(\mathbf{x}^{k+1}) - (\mathbf{c}^k + Q^k \mathbf{d}^k) + (\mathbf{c}^k + Q^k \mathbf{d}^k)\| \\ &\leq \|\mathbf{g}(\mathbf{x}^{k+1}) - (\mathbf{c}^k + Q^k \mathbf{d}^k)\| + \|(\mathbf{c}^k + Q^k \mathbf{d}^k)\| \\ &\leq \beta \|\mathbf{d}^k\|^2 + \lambda^k \|\mathbf{d}^k\| = \beta (\alpha^k)^2 + \lambda^k \alpha^k = \frac{\epsilon}{\beta} + \frac{\lambda^k \sqrt{\epsilon}}{\beta}. \end{aligned}$$

Thus, one can stop the algorithm as soon as $\lambda^k = \sqrt{\epsilon}$ so that the inequality becomes $\|\mathbf{g}(\mathbf{x}^{k+1})\| \leq \frac{2\epsilon}{\beta}$ and the function value is decreased at least $-\frac{\epsilon^{1.5}}{6\beta^2}$. Furthermore, $|\lambda_{\min}(\nabla \mathbf{g}(\mathbf{x}^k))| \leq \lambda^k = \sqrt{\epsilon}$.

Theorem 5 *Let the objective function $p^* = \inf f(\mathbf{x})$ be finite. Then in $\frac{O(\beta^2(f(\mathbf{x}^0) - p^*))}{\epsilon^{1.5}}$ iterations of the second-order method, the norm of the gradient vector is less than ϵ and the Hessian is $\sqrt{\epsilon}$ -positive semidefinite, where each iteration solves a spherical-constrained quadratic minimization discussed earlier.*

Adaptive Spherical Trust-Region Method

One can treat α as a variable in

$$\begin{aligned} \mathbf{d}^k = & \arg \min_{(\mathbf{d}, \alpha)} \quad (\mathbf{c}^k)^T \mathbf{d} + \frac{1}{2} \mathbf{d}^T Q^k \mathbf{d} + \frac{\beta}{3} \alpha^3 \\ \text{s.t.} \quad & \|\mathbf{d}\| \leq \alpha. \end{aligned}$$

Then, the optimality conditions of this sub-problem would be

$$(Q^k + \lambda I) \mathbf{d}^k = -\mathbf{c}^k, \quad (Q^k + \lambda I) \succeq \mathbf{0}, \quad \|\mathbf{d}\|_2^2 = \alpha^2,$$

and $\alpha = \frac{\lambda}{\beta}$. Thus, let $\mathbf{d}(\lambda) = -(Q^k + \lambda I)^{-1} \mathbf{c}^k$, the problem becomes finding the root λ of

$$\|\mathbf{d}(\lambda)\|^2 - \frac{\lambda^2}{\beta^2} = 0,$$

where $\lambda \geq -\lambda_{\min}(Q^k) > 0$ (assume that the current Hessian is not PSD yet), as in the Hybrid of Bisection and Newton method discussed earlier in $\log \log(1/\epsilon)$ arithmetic operations.

In practice, even β is unknown, one can forward/backward choose λ such as the objective function is reduced by a sufficient quantity, and there is no need to find the exact root.

Relation to Quadratic Regularization/Proximal-Point Method

One can also interpret the Spherical Trust-Region method as the Quadratic Regularization Method

$$\mathbf{d}^k(\lambda^k) = \arg \min_{\mathbf{d}} \quad (\mathbf{c}^k)^T \mathbf{d} + \frac{1}{2} \mathbf{d}^T Q^k \mathbf{d} + \frac{1}{2} \lambda^k \|\mathbf{d}\|^2$$

where parameter λ^k makes $(Q^k + \lambda^k I) \succeq \mathbf{0}$, and $\lambda^k \rightarrow 0$.

Then consider the one-variable function

$$\phi(\lambda) := f(\mathbf{x}^k + \mathbf{d}^k(\lambda))$$

and do one variable minimization of $\phi(\lambda)$ over λ such that $\lambda \geq \max\{0, -\lambda_{\min}(Q^k)\}$.

Let $\lambda^k = \arg \min \phi(\lambda)$ and $\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{d}^k(\lambda^k)$. Then, based on the above analysis, we must have at least

$$f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k) \leq -\frac{\epsilon^{1.5}}{6\beta^2}$$

where β is the (global) **Lipschitz** constant of the objective function. Note that the algorithm needs to estimate only the minimum eigenvalue of the Hessian.

Dimension-Reduced Second-Order Method with Trust Region

Let $\mathbf{d}^k = \mathbf{x}^k - \mathbf{x}^{k-1}$, $\mathbf{g}^k = \nabla f(\mathbf{x}^k)$ and $H^k = \nabla^2 f(\mathbf{x}^k)$, and

$$Q^k = \begin{pmatrix} (\mathbf{g}^k)^T H^k \mathbf{g}^k & -(\mathbf{d}^k)^T H^k \mathbf{g}^k \\ -(\mathbf{d}^k)^T H^k \mathbf{g}^k & (\mathbf{d}^k)^T H^k \mathbf{d}^k \end{pmatrix} \in S^2, \mathbf{c}^k = \begin{pmatrix} -\|\mathbf{g}^k\|^2 \\ (\mathbf{g}^k)^T \mathbf{d}^k \end{pmatrix} \in R^2.$$

Then, similar to the full-dimensional Spherical Trust-Region, one can construct a 2-dimensional trust-region quadratic model:

$$\alpha^k(\lambda^k) = \arg \min_{\alpha \in R^2} (\mathbf{c}^k)^T \alpha + \frac{1}{2} \alpha^T Q^k \alpha + \frac{1}{2} \lambda^k \|\alpha\|^2$$

where parameter λ^k makes $(Q^k + \lambda^k I) \succeq \mathbf{0}$. Finally let

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_1^k \mathbf{g}^k + \alpha_2^k \mathbf{d}^k.$$

Again, if the Hessian $\nabla^2 f(\mathbf{x}^k)$ is not available, one can approximate

$$H^k \mathbf{g}^k \sim \nabla(\mathbf{x}^k + \mathbf{g}^k) - \mathbf{g}^k \quad \text{and} \quad H^k \mathbf{d}^k \sim \nabla(\mathbf{x}^k + \mathbf{d}^k) - \mathbf{g}^k \sim -(\mathbf{g}^{k-1} - \mathbf{g}^k);$$

or more accurate difference approximation between two gradients.

Would Convexity Help?

Before we answer this question, let's summarize a generic form one iteration of the Second Order Method for solving $\nabla f(\mathbf{x}) = \mathbf{g}(\mathbf{x}) = \mathbf{0}$:

$$\begin{aligned} (\nabla \mathbf{g}(\mathbf{x}^k) + \mu I)(\mathbf{x} - \mathbf{x}^k) &= -\gamma \mathbf{g}(\mathbf{x}^k), \quad \text{or} \\ \mathbf{g}(\mathbf{x}^k) + \nabla \mathbf{g}(\mathbf{x}^k)(\mathbf{x} - \mathbf{x}^k) + \mu(\mathbf{x} - \mathbf{x}^k) &= (1 - \gamma)\mathbf{g}(\mathbf{x}^k). \end{aligned}$$

Many interpretations: when

- $\gamma = 1, \mu = 0$: pure **Newton**;
- γ and μ are sufficiently large: **SDM**;
- $\gamma = 1$ and μ decreases to 0: **Homotopy or path-following** method.

A Path-Following Algorithm for Unconstrained Optimization I

For any $\mu > 0$ consider the (unique) optimal solution $\mathbf{x}(\mu)$ for problem

$$\mathbf{x}(\mu) = \arg \min \mathbf{f}(\mathbf{x}) + \frac{\mu}{2} \|\mathbf{x}\|^2,$$

and they form a **path** down to $\mathbf{x}(0)$ and satisfy gradient equations with parameter μ :

$$\mathbf{g}(\mathbf{x}) + \mu \mathbf{x} = \mathbf{0}, \quad \text{with } \mu = \mu^k > 0. \quad (6)$$

Let the approximation path error at \mathbf{x}^k with $\mu = \mu^k$ be

$$\|\mathbf{g}(\mathbf{x}^k) + \mu^k \mathbf{x}^k\| \leq \frac{1}{2\beta} \mu^k.$$

Then, we like to compute a new iterate \mathbf{x}^{k+1} , using Newton's method with \mathbf{x}^k as an initial solution, such that

$$\|\mathbf{g}(\mathbf{x}^{k+1}) + \mu^{k+1} \mathbf{x}^{k+1}\| \leq \frac{1}{2\beta} \mu^{k+1}, \quad \text{where } 0 \leq \mu^{k+1} < \mu^k.$$

If μ^k can be decreased at a **geometric** rate, independent of ϵ , and each update uses one Newton step, then this would lead to a **linearly convergent** algorithm.

Concordant Lipschitz Functions

We analyze the path-following algorithm when f is convex and meet a **Concordant Lipschitz** condition: for any point \mathbf{x} and a $\beta \geq 1$

$$\|\nabla f(\mathbf{x} + \mathbf{d}) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})\mathbf{d}\| \leq \beta \mathbf{d}^T \nabla^2 f(\mathbf{x})\mathbf{d}, \text{ whenever } \|\mathbf{d}\| \leq O(1) < 1 \quad (7)$$

and $\mathbf{x} + \mathbf{d}$ in the function domain. Such condition can be verified using Taylor Expansion Series; basically, the third derivative of the function is bounded by its second derivative.

- All quadratic functions are concordant Lipschitz with $\beta = 0$.
- Convex function e^x is concordant Lipschitz with $\beta = O(1)$ but it is not regular Lipschitz.
- Convex function $-\log(x)$ is neither regular Lipschitz nor concordant Lipschitz.
- Function $f(\mathbf{x}) := \phi(A\mathbf{x} - \mathbf{b})$ is concordant Lipschitz if $\phi(\cdot)$ is regular Lipschitz and strictly convex.

A Path-Following Algorithm for Unconstrained Optimization II

When μ^k is replaced by μ^{k+1} , say $(1 - \eta)\mu^k$ for some $\eta \in (0, 1]$, we aim to find a solution \mathbf{x} such that

$$\mathbf{g}(\mathbf{x}) + (1 - \eta)\mu^k \mathbf{x} = \mathbf{0},$$

we start from \mathbf{x}^k and apply the **Newton iteration**:

$$\begin{aligned} \mathbf{g}(\mathbf{x}^k) + \nabla \mathbf{g}(\mathbf{x}^k) \mathbf{d} + (1 - \eta)\mu^k (\mathbf{x}^k + \mathbf{d}) &= \mathbf{0}, \quad \text{or} \\ \nabla \mathbf{g}(\mathbf{x}^k) \mathbf{d} + (1 - \eta)\mu^k \mathbf{d} &= -\mathbf{g}(\mathbf{x}^k) - (1 - \eta)\mu^k \mathbf{x}^k. \end{aligned} \tag{8}$$

From the second expression, we have

$$\begin{aligned} \|\nabla \mathbf{g}(\mathbf{x}^k) \mathbf{d} + (1 - \eta)\mu^k \mathbf{d}\| &= \|-\mathbf{g}(\mathbf{x}^k) - (1 - \eta)\mu^k \mathbf{x}^k\| \\ &= \|-\mathbf{g}(\mathbf{x}^k) - \mu^k \mathbf{x}^k + \eta\mu^k \mathbf{x}^k\| \\ &\leq \|-\mathbf{g}(\mathbf{x}^k) - \mu^k \mathbf{x}^k\| + \eta\mu^k \|\mathbf{x}^k\| \\ &\leq \frac{1}{2\beta} \mu^k + \eta\mu^k \|\mathbf{x}^k\|. \end{aligned} \tag{9}$$

On the other hand

$$\|\nabla \mathbf{g}(\mathbf{x}^k)\mathbf{d} + (1 - \eta)\mu^k\mathbf{d}\|^2 = \|\nabla \mathbf{g}(\mathbf{x}^k)\mathbf{d}\|^2 + 2(1 - \eta)\mu^k\mathbf{d}^T\nabla \mathbf{g}(\mathbf{x}^k)\mathbf{d} + ((1 - \eta)\mu^k)^2\|\mathbf{d}\|^2.$$

From **convexity**, $\mathbf{d}^T\nabla \mathbf{g}(\mathbf{x}^k)\mathbf{d} \geq 0$, together with (9) we have

$$\begin{aligned} ((1 - \eta)\mu^k)^2\|\mathbf{d}\|^2 &\leq \left(\frac{1}{2\beta} + \eta\|\mathbf{x}^k\|\right)^2(\mu^k)^2 \quad \text{and} \\ 2(1 - \eta)\mu^k\mathbf{d}^T\nabla \mathbf{g}(\mathbf{x}^k)\mathbf{d} &\leq \left(\frac{1}{2\beta} + \eta\|\mathbf{x}^k\|\right)^2(\mu^k)^2. \end{aligned}$$

The first inequality implies

$$\|\mathbf{d}\|^2 \leq \left(\frac{1}{2\beta(1 - \eta)} + \frac{\eta}{1 - \eta}\|\mathbf{x}^k\|\right)^2.$$

Let the new iterate be $\mathbf{x}^+ = \mathbf{x}^k + \mathbf{d}$. The second inequality implies

$$\begin{aligned} &\|\mathbf{g}(\mathbf{x}^+) + (1 - \eta)\mu^k\mathbf{x}^+\| \\ = &\|\mathbf{g}(\mathbf{x}^+) - (\mathbf{g}(\mathbf{x}^k) + \nabla \mathbf{g}(\mathbf{x}^k)\mathbf{d}) + (\mathbf{g}(\mathbf{x}^k) + \nabla \mathbf{g}(\mathbf{x}^k)\mathbf{d}) + (1 - \eta)\mu^k(\mathbf{x}^k + \mathbf{d})\| \\ = &\|\mathbf{g}(\mathbf{x}^+) - \mathbf{g}(\mathbf{x}^k) + \nabla \mathbf{g}(\mathbf{x}^k)\mathbf{d}\| \\ \leq &\beta\mathbf{d}^T\nabla \mathbf{g}(\mathbf{x}^k)\mathbf{d} \leq \frac{\beta}{2(1 - \eta)}\left(\frac{1}{2\beta} + \eta\|\mathbf{x}^k\|\right)^2\mu^k. \end{aligned}$$

We now just need to choose $\eta \in (0, 1)$ such that

$$\begin{aligned} \left(\frac{1}{2\beta(1-\eta)} + \frac{\eta}{1-\eta} \|\mathbf{x}^k\| \right)^2 &\leq 1 \quad \text{and} \\ \frac{\beta\mu^k}{2(1-\eta)} \left(\frac{1}{2\beta} + \eta \|\mathbf{x}^k\| \right)^2 &\leq \frac{1}{2\beta} (1-\eta) \mu^k = \frac{1}{2\beta} \mu^{k+1}. \end{aligned}$$

For example, given $\beta \geq 1$,

$$\eta = \frac{1}{2\beta(1 + \|\mathbf{x}^k\|)}$$

would suffice.

This would give a **linear convergence** since $\|\mathbf{x}^k\|$ is typically bounded following the path to the optimality, while the convergence in non-convex case is only arithmetic.

Convexity, together with some types of second-order methods, make convex optimization solvers into practical technologies.

A Path-Following Algorithm for Unconstrained Optimization III

More question related to the path-following algorithm:

- For convex case, since $\mathbf{x}(\mu)$ is the unique minimizer of

$$\min \mathbf{f}(\mathbf{x}) + \frac{\mu}{2} \|\mathbf{x}\|^2,$$

what is the limit of $\mathbf{x}(\mu)$ as $\mu \rightarrow 0^+$?

- More practical strategy to decrease μ ?
- Apply first-order or 1.5-order algorithms for solving each step of the path-following, since it is to minimize a strictly convex quadratic function?
- What happen when \mathbf{f} is bounded from below but not convex, and just meet the standard Lipschitz condition? The key is analyzing $\mathbf{x}(\mu)$, which may form multiple paths. Then can we still follow the path?