Quasi-Newton Methods

Lecturer: Aarti Singh
Co-instructor: Pradeep Ravikumar

Convex Optimization 10-725/36-725

Materials courtesy: B. Poczos, R. Tibshirani, J. Pena

Modified Newton Method

Goal: $$\min_{x \in \mathbb{R}^n} f(x)$$

Gradient descent:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \ \alpha_k > 0$$

Newton method:

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

Modified Newton method: [Method of Deflected Gradients]

$$x_{k+1} = x_k - \alpha_k S_k \nabla f(x_k)$$

$$S_k \in \mathbb{R}^{n \times n}, \ \alpha_k \in \mathbb{R}$$

Special cases: $S_k = I_n$: Gradient descent

$S_k = [\nabla^2 f(x_k)]^{-1}$: Newton method

Modified Newton Method

$$x_{k+1} = x_k - \alpha_k S_k \nabla f(x_k)$$

Lemma [Descent direction]

$S_k \succ 0 \Rightarrow$ the modified Newton step is a descent direction

Proof:

We know that if a vector has negative inner product with the gradient vector, then that direction is a descent direction

$$\Rightarrow \nabla f(x_k)^T (x_{k+1} - x_k) = -\nabla f(x_k)^T \alpha_k S_k \nabla f(x_k) < 0$$

Quadratic problem

$$\min_{x\in\mathbb{R}^n} f(x) \qquad f(x) = \frac{1}{2}x^T Q x - b^T x$$

Assume matrix $Q \in \mathbb{R}^{n\times n}$ is positive definite

Let $g_k \doteq \nabla f(x_k) = Q x_k - b$

Modified Newton Method update rule:

$$x_{k+1} = x_k - \alpha_k S_k g_k$$

Lemma [α_k in quadratic problems**]**

Let $\alpha_k = \arg\min_\alpha f(x_k - \alpha S_k g_k)$

$$\Rightarrow \alpha_k = \frac{g_k^T S_k g_k}{g_k^T S_k Q S_k g_k}$$

Quadratic problem

Lemma [α_k in quadratic problems]

$$f(x) = \frac{1}{2}x^T Q x - b^T x$$

$$g_k \doteq \nabla f(x_k) = Q x_k - b$$

Let $\alpha_k = \arg\min_\alpha f(x_k - \alpha S_k g_k)$

$$\Rightarrow \alpha_k = \frac{g_k^T S_k g_k}{g_k^T S_k Q S_k g_k}$$

Proof [α_k]

$$f(x) = \frac{1}{2}[x_k - \alpha S_k g_k]^T Q [x_k - \alpha S_k g_k] - b^T [x_k - \alpha S_k g_k]$$

$$0 = \nabla f(\alpha_k) = -g_k^T S_k Q [x_k - \alpha_k S_k g_k] + b^T S_k g_k$$

$$\Rightarrow \alpha_k g_k^T S_k Q S_k g_k = \underbrace{g_k^T S_k Q x_k - g_k^T S_k b}_{g_k^T S_k g_k}$$

$$\Rightarrow \alpha_k = \frac{g_k^T S_k g_k}{g_k^T S_k Q S_k g_k}$$

Convergence rate (Quadratic case)

Theorem [Convergence rate of the modified Newton method**]**

Let x^* be the unique minimum point of f.

Let $\epsilon(x_k) = \frac{1}{2}(x_k - x^*)^T Q(x_k - x^*)$ [Error of x_k]

Then for the modified Newton method it holds at every step k

$$\epsilon(x_{k+1}) \le \left(\frac{B_k - b_k}{B_k + b_k}\right)^2 \epsilon(x_k)$$

where b_k and B_k are, respectively, the smallest and largest eigenvalues of the matrix $S_k Q$

Superlinear in general

Linear if $S_k = I_n$: Gradient descent

Quadratic if $S_k = [\nabla^2 f(x_k)]^{-1}$: Newton method

Quasi-Newton Methods

Quasi-Newton Methods

Two main steps in Newton's method:

- Compute Hessian $\nabla^2 f(x)$
- Solve the system of equations

$$\nabla^2 f(x) p = -\nabla f(x).$$

Each of these two steps could be expensive.

Quasi-Newton method

Use instead

$$x^+ = x + tp$$

where

$$Bp = -\nabla f(x)$$

for some approximation B of $\nabla^2 f(x)$.

Want B easy to compute and $Bp = g$ easy to solve.

Secant Equation

We would like B^k to approximate $\nabla^2 f(x^k)$, that is

$$\nabla f(x^k + s) \approx \nabla f(x^k) + B^k s.$$

Once $x^{k+1} = x^k + s^k$ is computed, we would like a new B^{k+1}.

Idea: since B^k already contains some information, make some suitable update.

Reasonable requirement for B^{k+1}

$$\nabla f(x^{k+1}) = \nabla f(x^k) + B^{k+1} s^k$$

or equivalently

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

Secant Equation

The latter condition is called the secant equation and written as

$$B^{k+1}s^k = y^k \text{ or simply } B^+s = y$$

where $s^k = x^{k+1} - x^k$ and $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$.

In addition to the secant equation, we would like

(i) B^+ symmetric

(ii) B^+ "close" to B

(iii) B positive definite $\Rightarrow B^+$ positive definite

Symmetric Rank-1 (SR1) method

Symmetric Rank-1 (SR1) update

Try an update of the form

$$B^+ = B + auu^\mathsf{T}$$

Let $H = B^{-1}$. Try updating the inverse Hessian directly.

$$H^+ = H + bzz^\top$$

We will derive the following SR1 updates that satisfy the secant equation:

$$H^+ = H + \frac{(s - Hy)(s - Hy)^\mathsf{T}}{(s - Hy)^\mathsf{T} y}$$

$$B^+ = B + \frac{(y - Bs)(y - Bs)^\mathsf{T}}{(y - Bs)^\mathsf{T} s}$$

Symmetric Rank-1 (SR1) update

Secant equation implies

$$B^{+}s = y \;\Rightarrow\; s = H^{+}y$$

So we get

$$s = (H + bzz^{\top})y = Hy + bzz^{\top}y \qquad (1)$$

$$\Rightarrow s - Hy = bzz^{\top}y$$

$$\Rightarrow \frac{(s-Hy)(s-Hy)^{\top}}{b} = bzz^{\top}yy^{\top}zz^{\top} = bz(z^{\top}y)^{2}z^{\top}$$

$$\Rightarrow \boxed{bzz^{\top} = \frac{(s-Hy)(s-Hy)^{\top}}{b(z^{\top}y)^{2}}}$$

Also from (1) $\quad y^{\top}s = y^{\top}Hy + b(z^{\top}y)^{2}$

$$\Rightarrow \boxed{b(z^{\top}y)^{2} = y^{\top}(s - Hy)}$$

Symmetric Rank-1 (SR1) update

SR1 update for inverse Hessian $H^+ = H + bzz^\top$

$$H^+ = H + \frac{(s - Hy)(s - Hy)^\mathsf{T}}{(s - Hy)^\mathsf{T} y}$$

A low-rank update on a matrix corresponds to a low rank update on its inverse.

Theorem (Sherman-Morrison-Woodbury formula)

Assume $A \in \mathbb{R}^{n \times n}$, and $U, V \in \mathbb{R}^{n \times d}$. Then $A + UV^\mathsf{T}$ is nonsingular if and only if $I + V^\mathsf{T} A^{-1} U$ is nonsingular. In that case

$$(A + UV^\mathsf{T})^{-1} = A^{-1} - A^{-1}U(I + V^\mathsf{T} A^{-1} U)^{-1} V^\mathsf{T} A^{-1}$$

SR1 update for Hessian

$$B^+ = B + \frac{(y - Bs)(y - Bs)^\mathsf{T}}{(y - Bs)^\mathsf{T} s}$$

Symmetric Rank-1 (SR1) update

Algorithm: [Modified Newton method with rank 1 correction]

$$x_{k+1} = x_k - \alpha_k H_k g_k$$

where $\alpha_k = \arg\min_\alpha f(x_k - \alpha H_k g_k)$ [Line search]

$$g_k = \nabla f(x_k)$$

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^\top}{(s_k - H_k y_k)^\top y_k}$$

$$s_k = x_{k+1} - x_k \qquad y_k = g_{k+1} - g_k$$

Issues:

Although H_k is symmetric, it might not be positive definite.

If $(s_k - H_k y_k)^\top y_k$ close to zero, then it is numerically unstable.

Davidon-Fletcher-Powell (DFP) update

Davidon-Fletcher-Powell (DFP) update

Try a rank-two update

$$H^+ = H + auu^\mathsf{T} + bvv^\mathsf{T}.$$

The secant equation yields

$$s - Hy = (au^\mathsf{T}y)u + (bv^\mathsf{T}y)v.$$

Putting $u = s,\ v = Hy$, and solving for a, b we get

$$H^+ = H - \frac{Hyy^\mathsf{T}H}{y^\mathsf{T}Hy} + \frac{ss^\mathsf{T}}{y^\mathsf{T}s}$$

By Sherman-Morrison-Woodbury we get a rank-two update on B

$$\begin{aligned}B^+ &= B + \frac{(y - Bs)y^\mathsf{T}}{y^\mathsf{T}s} + \frac{y(y - Bs)^\mathsf{T}}{y^\mathsf{T}s} - \frac{(y - Bs)^\mathsf{T}s}{(y^\mathsf{T}s)^2}yy^\mathsf{T} \\ &= \left(I - \frac{ys^\mathsf{T}}{y^\mathsf{T}s}\right) B \left(I - \frac{sy^\mathsf{T}}{y^\mathsf{T}s}\right) + \frac{yy^\mathsf{T}}{y^\mathsf{T}s}\end{aligned}$$

DFP method

$H_0 \in \mathbb{R}^{n\times n}$ initial symmetric, pos. def. matrix.

$x_0 \in \mathbb{R}^n$, $k = 0$ $\quad g_k = \nabla f(x_k)$

Step 1. $d_k = -H_k g_k$ [Search direction]

Step 2. $\alpha_k = \arg\min_{\alpha>0} f(x_k + \alpha d_k)$ [Line search]

$$x_{k+1} = x_k + \alpha_k d_k$$

$$s_k = x_{k+1} - x_k = \alpha_k d_k$$

$$g_{k+1} = \nabla f(x_{k+1})$$

Step 3. $y_k = g_{k+1} - g_k$

$$H_{k+1} = H_k - \frac{H_k y_k y_k^\top H_k}{y_k^\top H_k y_k} + \frac{s_k s_k^\top}{y_k^\top s_k}$$ [rank 2 update]

$k = k + 1$ and return to Step 1.

DFP method

Theorem [H_k is positive definite**]**

In the DFP method if $H_0 \succ 0$, then $H_k \succ 0$.

Theorem [DFP is a conjugate direction method**]**

If f is quadratic with positive definite Hessian Q, then

$$d_i^\top Q d_j = 0, \; 0 \leq i < j \leq k$$

Corollary [finite step convergence for quadratic functions**]**

If f is quadratic with positive definite Hessian Q, then $H_n = Q^{-1}$

DFP update – alternate derivation

Find B^+ closest to B in some norm so that B^+ satisfies the secant equation and is symmetric:

$$\begin{aligned} \min_{B^+} \quad & \|B^+ - B\|_? \\ \text{subject to} \quad & B^+ = (B^+)^\mathsf{T} \\ & B^+ s = y \end{aligned}$$

What norm to use?

DFP update – alternate derivation

Observe: B^+ positive definite and $B^+s = y$ imply

$$y^\mathsf{T} s = s^\mathsf{T} B^+ s > 0.$$

The inequality $y^\mathsf{T} s > 0$ is called the curvature condition.

Fact: if $y, s \in \mathbb{R}^n$ and $y^\mathsf{T} s > 0$ then there exists M symmetric and positive definite such that $Ms = y$.

DFP update again

Solve

$$\begin{array}{rl} \displaystyle\min_{B^+} & \|W^{-1}(B^+ - B)W^{-\mathsf{T}}\|_F \\ \text{subject to} & B^+ = (B^+)^\mathsf{T} \\ & B^+ s = y \end{array}$$

where $W \in \mathbb{R}^{n \times n}$ is nonsingular and such that $WW^\mathsf{T} s = y$.

Broyden-Fletcher-Goldfarb-Shanno (BFGS) method

Broyden-Fletcher-Goldfarb-Shanno (BFGS) update

Same ideas as the DFP update but with roles of B and H exchanged.

Closeness to H:

$$\begin{array}{ll} \min\limits_{H^+} & \|W^{-1}(H^+ - H)W^{-\mathsf{T}}\|_F \\ \text{subject to} & H^+ = (H^+)^\mathsf{T} \\ & H^+ y = s \end{array}$$

where $W \in \mathbb{R}^{n \times n}$ is nonsingular and $WW^\mathsf{T} y = s$.

BFGS update

Swapping H and B and y and s in the DFP update we get

$$B^+ = B - \frac{Bss^\mathsf{T}B}{s^\mathsf{T}Bs} + \frac{yy^\mathsf{T}}{y^\mathsf{T}s}$$

and

$$\begin{aligned} H^+ &= H + \frac{(s - Hy)s^\mathsf{T}}{y^\mathsf{T}s} + \frac{s(s - Hy)^\mathsf{T}}{y^\mathsf{T}s} - \frac{(s - Hy)^\mathsf{T}y}{(y^\mathsf{T}s)^2}ss^\mathsf{T} \\ &= \left(I - \frac{sy^\mathsf{T}}{y^\mathsf{T}s}\right) H \left(I - \frac{ys^\mathsf{T}}{y^\mathsf{T}s}\right) + \frac{ss^\mathsf{T}}{y^\mathsf{T}s} \end{aligned}$$

Both DFP and BFGS preserve positive definiteness: if B is positive definite and $y^\mathsf{T}s > 0$ then B^+ is positive definite.

In practice BFGS seems to work better than DFP.