Convex Optimization
CMU-10725

Quasi Newton Methods

Barnabás Póczos & Ryan Tibshirani

Quasi Newton Methods

Outline

- Modified Newton Method
- Rank one correction of the inverse
- Rank two correction of the inverse
  - Davidon–Fletcher–Powell Method (DFP)
  - Broyden–Fletcher–Goldfarb–Shanno Method (BFGS)

Books to Read

David G. Luenberger, Yinyu Ye: Linear and Nonlinear Programming

Nesterov: Introductory lectures on convex optimization

Bazaraa, Sherali, Shetty: Nonlinear Programming

Dimitri P. Bestsekas: Nonlinear Programming

Motivation

Motivation:

Evaluation and use of the Hessian matrix is impractical or costly

Idea:

use an approximation to the inverse Hessian.

Quasi Newton:

somewhere between steepest descent and Newton's method

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:

$$S_k = I_n: \text{ Gradient descent}$$

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

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]

x_{k+1}

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

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

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

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

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

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

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

Corollary

If S_k^{-1} is close to Q, then b_k is close to B_k, and then convergence is fast

Proof: No time for it…

Classical modified Newton's method

Classical modified Newton:

Standard method for approximating the Hessian without evaluating $[\nabla^2 f(x_k)]^{-1}$ for each k.

$$x_{k+1} = x_k - \alpha_k [\nabla^2 f(x_0)]^{-1} \nabla f(x_k)$$

The Hessian at the initial point x_0 is used throughout the process.

The effectiveness depends on how fast the Hessian is changing.

Construction of the inverse of the Hessian

Construction of the inverse

Idea behind quasi-Newton methods: construct the approximation of the inverse Hessian using information gathered during the process

We show how the inverse Hessian can be built up from gradient information obtained at various points.

Notation:

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

$$p_k = x_{k+1} - x_k \qquad Q(x_k) = \nabla^2 f(x_k)$$

$$q_k = g_{k+1} - g_k = \nabla f(x_{k+1}) - \nabla f(x_k) \approx Q(x_k) p_k$$

In the quadratic case

$g_k \doteq \nabla f(x_k) = Qx_k - b$, and therefore

$$q_k = g_{k+1} - g_k = Q(x_{k+1} - x_k) = Qp_k$$

Construction of the inverse

Quadratic case: $q_k = Qp_k$ Let $H = Q^{-1}$

If n linearly independent directions $p_0, p_1 \dots p_{n-1}$ and the corresponding $q_0, q_1, \dots q_{n-1}$ are known, then Q is uniquely determined.

$$[q_0, q_1, \dots q_{n-1}] = Q[p_0, p_1, \dots p_{n-1}]$$

$$\Rightarrow Q = [q_0, q_1, \dots q_{n-1}][p_0, p_1, \dots p_{n-1}]^{-1}$$

$$\Rightarrow H[q_0, q_1, \dots q_{n-1}] = [p_0, p_1, \dots p_{n-1}]$$

Goal:

We will construct successive approximations H_k to H based on data obtained from the first k steps such that

$$H_{k+1}[q_0, q_1, \dots q_k] = [p_0, p_1, \dots p_k]$$

After n linearly independent steps we would then have $H_n = H$.

Symmetric rank one correction (SR1)

We want an update on H_k **such that :**

$$H_{k+1}[q_0, q_1, \ldots q_k] = [p_0, p_1, \ldots p_k]$$

Let us find the update in this form [Rank one correction]

$$H_{k+1} = H_k + a_k z_k z_k^T$$

We need a good $a_k \in \mathbb{R}$ and $z_k \in \mathbb{R}^n$

Theorem [Rank one update of H_k**]**

If $H_{k+1}[q_0, q_1, \ldots q_k] = [p_0, p_1, \ldots p_k]$

and $H_{k+1} = H_k + a_k z_k z_k^T$

$$\Rightarrow H_{k+1} = H_k + \frac{(p_k - H_k q_k)(p_k - H_k q_k)^T}{q_k^T (p_k - H_k q_k)}$$

Symmetric rank one correction (SR1)

Proof: We already know that

$$[p_0, p_1, \ldots p_k] = H_{k+1}[q_0, q_1, \ldots q_k], \text{ and } H_{k+1} = H_k + a_k z_k z_k^T$$

Therefore,

$$p_k = H_{k+1} q_k = [H_k + a_k z_k z_k^T] q_k = H_k q_k + a_k z_k z_k^T q_k$$

$$p_k - H_k q_k = a_k z_k z_k^T q_k$$

$$\frac{(p_k - H_k q_k)(p_k - H_k q_k)^T}{a_k} = a_k z_k z_k^T q_k q_k^T z_k z_k^T = a_k z_k (z_k^T q_k)^2 z_k^T$$

$$H_{k+1} = H_k + \frac{(p_k - H_k q_k)(p_k - H_k q_k)^T}{a_k (z_k^T q_k)^2}$$

$$q_k^T p_k = q_k^T H_k q_k + a_k q_k^T z_k z_k^T q_k = q_k^T H_k q_k + a_k (q_k^T z_k)^2$$

$$\Rightarrow H_{k+1} = H_k + \frac{(p_k - H_k q_k)(p_k - H_k q_k)^T}{q_k^T (p_k - H_k q_k)}$$

Q.E.D.

Symmetric rank one correction (SR1)

We still have to proof that this update will be good for us:

Theorem [H_k update works**]**

Let $Q \in \mathbb{R}^n$ be a given positive definite matrix.

$p_i \in \mathbb{R}^n$ $(0 \leq i \leq k)$ given vectors.

- $q_i = Qp_i, \; \forall i = 0, 1, \ldots, k$

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

- If $H_{i+1} = H_i + \frac{(p_i - H_i q_i)(p_i - H_i q_i)^T}{q_i^T (p_i - H_i q_i)}$, then

$$H_{k+1}[q_0, q_1, \ldots q_k] = [p_0, p_1, \ldots p_k]$$

Corollary

If $p_0, \ldots p_{n-1}$ are independent $\Rightarrow H_n = H = Q^{-1}$.

Symmetric rank one correction (SR1)

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]

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

$$H_{k+1} = H_k + \frac{(p_k - H_k q_k)(p_k - H_k q_k)^T}{q_k^T (p_k - H_k q_k)}$$

$$p_k = x_{k+1} - x_k \qquad q_k = g_{k+1} - g_k$$

Issues:

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

If $q_k^T (p_k - H_k q_k)$ is close to zero, then it is numerically unstable.

Davidon–Fletcher–Powell Method [Rank two correction]

Davidon–Fletcher–Powell Method

- For a quadratic objective, it simultaneously generates the directions of the conjugate gradient method while constructing the inverse Hessian.

- At each step the inverse Hessian is updated by the sum of **two** symmetric rank one matrices. **[rank two correction procedure]**

- The method is also often referred to as the variable metric method

Davidon–Fletcher–Powell 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]

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

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

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

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

$$H_{k+1} = H_k + \frac{p_k p_k^T}{p_k^T q_k} - \frac{H_k q_k q_k^T H_k}{q_k^T H_k q_k},$$ [rank 2 update]

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

Davidon–Fletcher–Powell 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 for the Davidon-Fletcher-Powell method

$$p_i^T Q p_j = 0, \; 0 \leq i < j \leq k$$

$$H_{k+1} Q p_i = p_i, \; 0 \leq i \leq k$$

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

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

Broyden–Fletcher–Goldfarb–Shanno

In DFP, at each step the inverse Hessian is updated by the sum of two symmetric rank one matrices.

BFGS we will estimate the Hessian Q, instead of its inverse

In the quadratic case we already proved:

$$Q[p_0, p_1, \ldots p_{n-1}] = [q_0, q_1, \ldots q_{n-1}]$$

$$H[q_0, q_1, \ldots q_{n-1}] = [p_0, p_1, \ldots p_{n-1}]$$

To estimate H, we used the update:

$$H_{k+1} = H_k + \frac{(p_k - H_k q_k)(p_k - H_k q_k)^T}{q_k^T (p_k - H_k q_k)}$$

Therefore, if we switch q and p, then Q can be estimated as well with Q_k

$$Q_{k+1} = Q_k + \frac{(q_k - Q_k p_k)(q_k - Q_k p_k)^T}{p_k^T (q_k - Q_k p_k)}$$

BFGS

Similarly, we already know that the the DFP update rule for H is

$$H_{k+1}^{DFP} = H_k + \frac{p_k p_k^T}{p_k^T q_k} - \frac{H_k q_k q_k^T H_k}{q_k^T H_k q_k},$$

Switching q and p, this can also be used to estimate Q:

$$Q_{k+1}^{BFGS} = Q_k + \frac{q_k q_k^T}{q_k^T p_k} - \frac{Q_k p_k p_k^T Q_k}{p_k^T Q_k p_k},$$

In the minimization algorithm, however, we will need an estimator of Q^{-1}

Let $H_{k+1}^{BFGS} \doteq Q_{k+1}^{-1}$

To get an update for H_{k+1}, let us use the Sherman-Morrison formula twice

$$H_{k+1}^{BFGS} = H_k^{BFGS} + \left(1 + \frac{q_k^T H_k^{BFGS} q_k}{p_k^T q_k}\right) \frac{p_k p_k^T}{p_k^T q_k} - \frac{p_k q_k^T H_k^{BFGS} + H_k^{BFGS} q_k p_k^T}{q_k^T p_k}$$

Sherman-Morrison matrix inversion formula

Suppose A is an invertible square matrix and u, v are vectors.

Suppose furthermore that $1 + v^T A^{-1} u \neq 0$.

Then the Sherman-Morrison formula states that

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} uv^T A^{-1}}{1 + v^T A^{-1} u}$$

BFGS Algorithm

$H_0 \in \mathbb{R}^{n\times n}$ 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]

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

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

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

$$H_{k+1}^{BFGS} = H_k^{BFGS} + \left(1 + \frac{q_k^T H_k^{BFGS} q_k}{p_k^T q_k}\right)\frac{p_k p_k^T}{p_k^T q_k} - \frac{p_k q_k^T H_k^{BFGS} + H_k^{BFGS} q_k p_k^T}{q_k^T p_k}$$

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

BFGS is almost the same as DFP, only the H update is different.

In practice BFGS seems to work better than DFP.

Summary

- Modified Newton Method
- Rank one correction of the inverse
- Rank two correction of the inverse
  - Davidon–Fletcher–Powell Method (DFP)
  - Broyden–Fletcher–Goldfarb–Shanno Method (BFGS)