

# Optimization, Written Assignment #3

April 29, 2008

## 1 Conjugate Direction Methods

Recall that  $v^{(1)}, \dots, v^{(k)} \in \mathbb{R}^n$  are  $Q$ -conjugate iff for every  $i \neq j$  we have  $(v^{(i)})^T Q v^{(j)} = 0$ . That is,  $\tilde{v}^{(i)} = Q^{1/2} v^{(i)}$  are orthogonal.

### 1.1 Conjugate Direction Minimization of a Quadratic Objective

Let  $f(x) = \frac{1}{2}x^T Q x + b^T x$ , with  $Q$  positive semi-definite, be a convex quadratic objective. Let  $\Delta x^{(0)}, \dots, \Delta x^{(n-1)}$  be non-zero  $Q$ -conjugate directions. Consider iterative minimization along these directions, starting from some  $x^{(0)}$ :

1. For  $i = 0$  to  $n - 1$
2.  $\alpha^{(i)} \leftarrow \arg \min_{\alpha} f(x^{(i)} + \alpha \Delta x^{(i)})$
3.  $x^{(i+1)} \leftarrow x^{(i)} + \alpha^{(i)} \Delta x^{(i)}$

#### 1.1.1

Prove that:

$$\alpha^{(i)} = -\frac{(\Delta x^{(i)})^T (b + Qx^{(i)})}{(\Delta x^{(i)})^T Q \Delta x^{(i)}}$$

#### 1.1.2

The principal result about conjugate directions is that the current point  $x^{(k)}$  at each step  $k$  of the method above minimizes the quadratic objective  $f(x)$  over the  $k$ -dimensional affine subspace spanned by  $\Delta x^{(0)}, \dots, \Delta x^{(k-1)}$ . That is:

$$x^{(k)} = \arg \min_{x \in M^k} f(x) \tag{1}$$

where

$$M^k = \left\{ x \mid x = x^0 + \sum_{i=0}^{k-1} \beta_i \Delta x^{(i)}, \beta_i \in \mathbb{R} \right\}$$

Prove equation (1):

1. Show that for all  $i < k$ :  $\nabla f(x^{(k)})^T \Delta x^{(i)} = \nabla f(x^{(i+1)})^T \Delta x^{(i)}$ . (Hint: write  $x^{(k)}$  in terms of  $x^{(i+1)}$ ,  $\alpha^{(i+1)}, \dots, \alpha^{(k-1)}$  and  $\Delta x^{(i+1)}, \dots, \Delta x^{(k-1)}$ )
2. Show that  $\nabla f(x^{(i+1)})^T \Delta x^{(i)} = 0$ . Conclude that  $\nabla f(x^{(k)})^T \Delta x^{(i)} = 0$  for  $i < k$ . (Hint: Consider the derivative of  $f(x^{(i)} + \alpha \Delta x^{(i)})$  with respect to  $\alpha$ .)
3. Prove equation (1) by considering the derivatives of  $x^0 + \sum_{i=0}^{k-1} \beta_i \Delta x^{(i)}$  with respect to  $\beta_i$ .

## 1.2 Generating Conjugate Directions

Let  $\Delta x^{(0)}, \dots, \Delta x^{(k-1)}$  be  $Q$ -conjugate and  $d$  a non-zero vector which is not spanned by  $\Delta x^{(0)}, \dots, \Delta x^{(k-1)}$ . Let

$$\Delta x^{(k)} = d - \sum_{i=0}^{k-1} \frac{d^T Q \Delta x^{(i)}}{(\Delta x^{(i)})^T Q \Delta x^{(i)}} \Delta x^{(i)} \quad (2)$$

### 1.2.1

Prove that  $\Delta x^{(0)}, \dots, \Delta x^{(k)}$  are  $Q$ -conjugate and that they span the same subspace as  $\Delta x^{(0)}, \dots, \Delta x^{(k-1)}, d$ .

## 1.3 The Conjugate Gradient Method for a Quadratic Function

In the conjugate gradient method for a quadratic function  $f(x) = \frac{1}{2}x'Qx + b'x$ , each iteration starts with the negative gradient  $d = -\nabla f(x)$  and applies equation (2) to obtain only the part of  $d$  that is conjugate to all previous directions:

1. For  $i = 0$  to  $n - 1$
2.  $d^{(i)} = -\nabla f(x^{(i)})$
3. If  $d^{(i)} = 0$  then terminate
4. Calculate  $\Delta x^{(i)}$  using equation (2)
5.  $\alpha^{(i)} = -\frac{(\Delta x^{(i)})^T (b + Qx^{(i)})}{(\Delta x^{(i)})^T Q \Delta x^{(i)}}$
6.  $x^{(i+1)} \leftarrow x^{(i)} + \alpha^{(i)} \Delta x^{(i)}$

### 1.3.1

Explain why after running the above method, if the method does not terminate early, then  $x^{(n)}$  is an optimal point. If the method does terminate early, the last iterate is an optimal point.

### 1.3.2

The key to the conjugate gradient method is that the calculation of the direction  $\Delta x^{(i)}$  can be greatly simplified. In particular, we have:

$$\Delta x^{(k)} = d^{(k)} + \beta^{(k)} \Delta x^{(k-1)} \quad (3)$$

with

$$\beta^{(k)} = \frac{(d^{(k)})^T d^{(k)}}{d^{(k-1)T} d^{(k-1)}} \quad (4)$$

Prove equation (3):

1. Prove that  $d^{(k)}$  is orthogonal to  $\Delta x^{(0)}, \dots, \Delta x^{(k-1)}$  and hence also to  $d^{(0)}, \dots, d^{(k-1)}$ . (Hint: Use the partial optimality property given in equation (1)).
2. Show that  $\alpha^{(i)} Q \Delta x^{(i)} = d^{(i)} - d^{(i+1)}$ . (Hint: expand the gradients and consider the update rule for  $x^{(i+1)}$ ).
3. Using the above relation and the orthogonality of  $d^{(0)}, \dots, d^{(k)}$ , evaluate  $(d^{(i)})^T Q \Delta x^{(j)}$  for  $j < i$ . (Hint: For all but one value of  $j$ , this will be zero).
4. Similarly, evaluate  $(\Delta x^{(j)})^T Q \Delta x^{(j)}$ .

5. Substitute the above two relations into equation (2) and obtain equation (3), with  $\beta^{(k)}$  expressed in terms of  $d^{(k)}$ ,  $d^{(k-1)}$  and  $\Delta x^{(k-1)}$ . Now, show that  $\beta^{(k)}$  can be calculated as in equation (4) by expanding  $\Delta x^{(k-1)}$  using equation (3), the orthogonality of  $d^{(k)}$  and  $d^{(k-1)}$  and the orthogonality of  $\Delta x^{(k-2)}$  and  $d^{(k)} - d^{(k-1)}$ .

This concludes the proof of equations (3) and (4). We will actually prefer a slightly different form of equation (4):

$$\beta^{(k)} = \frac{(d^{(k)})^T (d^{(k)} - d^{(k-1)})}{d^{(k-1)T} d^{(k-1)}} \quad (5)$$

6. Show that equation (5) is also valid and equivalent to equation (4) (when minimizing a quadratic function with exact line search).

Each iteration of the method therefore requires only vector-vector operations with computational cost  $O(n)$ , once the gradient has been computed. For a quadratic function, the most expensive operation is therefore computing the gradient which takes time  $O(n^2)$ .

## 2 Quasi-Newton Methods

All parts of this problem are optional. However, you are encouraged to go through the claims in order to review the description of the method and its properties discussed in class.

In quasi-Newton methods the descent direction is given by:

$$\Delta x^{(k)} = -D^{(k)} \nabla f(x^{(k)})$$

In the exact Newton method, the matrix  $D^{(k)}$  is the inverse Hessian. Quasi-Newton methods avoid calculating the Hessian and inverting it by updating an approximation of the inverse Hessian using the change in the gradients. For a quadratic function, the change in gradient is described by:

$$q^{(k)} = (\nabla^2 f) p^{(k)}$$

where  $p^{(k)} = x^{(k+1)} - x^{(k)}$  and  $q^{(k)} = \nabla f(x^{(k+1)}) - \nabla f(x^{(k)})$ . We therefore seek an approximation  $D^{(k)}$  to the inverse Hessian that approximately satisfies:

$$p^{(k)} \approx D q^{(k)}$$

The Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton method updates  $D^{(k)}$  by making the smallest change, under some specific weighted norm, that agrees with the latest change in the gradient:

$$D^{(k+1)} = \arg \min_{p^{(k)} = D q^{(k)}} \left\| W^{1/2} (D - D^{(k)}) W^{1/2} \right\|_F \quad (6)$$

where  $\|A\|_F = \sqrt{\sum_{ij} A_{ij}^2}$  is the Frobenius norm and  $W$  is any matrix such that  $q^{(k)} = W p^{(k)}$ .

### 2.1

Show that the solution of equation (6) is given by:

$$D^{(k+1)} = D^{(k)} + \frac{p^{(k)} (p^{(k)})^T}{(p^{(k)})^T p^{(k)}} - \frac{D^{(k)} q^{(k)} (q^{(k)})^T D^{(k)}}{(q^{(k)})^T D^{(k)} q^{(k)}} + \tau^{(k)} v^{(k)} (v^{(k)})^T \quad (7)$$

where  $\tau^{(k)} = (q^{(k)})^T D^{(k)} q^{(k)}$ , and:

$$v^{(k)} = \frac{p^{(k)}}{(p^{(k)})^T p^{(k)}} - \frac{D^{(k)} q^{(k)}}{\tau^{(k)}}$$

The BFGS method is therefore given by (ignoring the stopping condition):

1. Start from some  $x^{(0)}$  and an initial  $D^{(0)}$

2. For  $i \in \{0, 1, 2, \dots\}$
3.  $\Delta x^{(i)} \rightarrow -D^{(i)} \nabla f(x^{(i)})$
4.  $\alpha^{(i)} \leftarrow \arg \min_{\alpha} f(x^{(i)} + \alpha \Delta x^{(i)})$
5.  $x^{(i+1)} \leftarrow x^{(i)} + \alpha^{(i)} \Delta x^{(i)}$
6. Calculate  $D^{(i+1)}$  according to equation (7)

## 2.2

We now consider applying BFGS to a quadratic objective  $f(x) = \frac{1}{2}x'Qx + b'x$  with  $x \in \mathbb{R}^n$  and  $Q$  positive definite.

### 2.2.1

Show that for all  $i < k \leq n$  we have  $D^{(k)}q^{(i)} = p^{(i)}$ . That is, for a quadratic objective, the approximate inverse Hessian matches all the changes in the gradient so far. Conclude that  $D^{(n)} = Q^{-1}$ , i.e. after  $n$  iterations the correct Hessian is recovered.

### 2.2.2

Show that  $\Delta x^{(0)}, \dots, \Delta x^{(n-1)}$  are  $Q$ -conjugate.

### 2.2.3

Show that with  $D^{(0)} = I$ , the sequence of iterates  $x^{(i)}$  generated by BFGS is identical to those generated by the conjugate gradient method described above. It is important to note that this holds only for a quadratic objective, and when exact line search is used. For non-quadratic objectives, or when approximate line search is used, the two methods typically differ.