Convex Optimization

Optional Enrichment Problem Set $1\frac{1}{2}$

Please do not turn this in—it will not be graded

1 Conjugate Direction Methods

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

1.1 Conjugate Direction Minimization of a Quadratic Objective

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

1. For $i = 0$ to $n-1$
2. $\quad t^{(i)} \leftarrow \arg\min_t f\left(x^{(i)} + t\Delta x^{(i)}\right)$
3. $\quad x^{(i+1)} \leftarrow x^{(i)} + t^{(i)} \Delta x^{(i)}$

1.1.1

Prove that:

$$t^{(i)} = \frac{\left(\Delta x^{(i)}\right)^T \left(H x^{(i)} - b\right)}{\left(\Delta x^{(i)}\right)^T H \Delta x^{(i)}}$$

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 $\Delta x^{(0)}, \ldots, \Delta x^{(k-1)}$. 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 $i < k$: $\nabla f\left(x^{(k)}\right)^T \Delta x^{(i)} = \nabla f\left(x^{(i+1)}\right)^T \Delta x^{(i)}$. (Hint: write $x^{(k)}$ in terms of $x^{(i+1)}$, $t^{(i+1)}, \dots, t^{(k-1)}$ and $\Delta x^{(i+1)}, \dots, \Delta x^{(k-1)}$)

2. Show that $\nabla f\left(x^{(i+1)}\right)^T \Delta x^{(i)} = 0$. Conclude that $\nabla f\left(x^{(k)}\right)^T \Delta x^{(i)} = 0$ for $i < k$. (Hint: Consider the derivative of $f\left(x^{(i)} + t\Delta x^{(i)}\right)$ with respect to t).

3. Prove equation (1) by considering the derivatives of $x^0 + \sum_{i=0}^{k-1} \beta_i \Delta x^{(i)}$ with respect to β_i.

1.2 Generating Conjugate Directions

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

$$\Delta x^{(k)} = d - \sum_{i=0}^{k-1} \frac{d^T H \Delta x^{(i)}}{\left(\Delta x^{(i)}\right)^T H \Delta x^{(i)}} \Delta x^{(i)} \tag{2}$$

1.2.1

Prove that $\Delta x^{(0)}, \dots, \Delta x^{(k)}$ are H-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 $f(x) = \frac{1}{2}x'Hx - b'x$, 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. $\quad d^{(i)} = -\nabla f\left(x^{(i)}\right)$
3. $\quad$ If $d^{(i)} = 0$ then terminate
4. $\quad$ Calculate $\Delta x^{(i)}$ using equation (2)
5. $\quad t^{(i)} = \frac{\left(\Delta x^{(i)}\right)^T \left(Hx^{(i)} - b\right)}{\left(\Delta x^{(i)}\right)^T H \Delta x^{(i)}}$
6. $\quad x^{(i+1)} \leftarrow x^{(i)} + t^{(i)} \Delta x^{(i)}$

1.3.1

Explain why after running the above method, if the method does not terminate early, than $x^{(n)}$ 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:

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

with

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

Prove equation (3):

1. Prove that $d^{(k)}$ is orthogonal to $\Delta x^{(0)}, \ldots, \Delta x^{(k-1)}$ and hence also to $d^{(0)}, \ldots, d^{(k-1)}$. (Hint: Use the partial optimality property given in equation (1)).

2. Show that $t^{(i)} H \Delta x^{(i)} = d^{(i)} - d^{(i+1)}$. (Hint: expand the gradients and consider the update rule for $x^{(i+1)}$).

3. Using the above relation and the orthogonality of $d^{(0)}, \ldots, d^{(k)}$, evaluate $\left(d^{(i)}\right)^T H \Delta x^{(j)}$ for $j < i$. (Hint: For all but one value of j, this will be zero).

4. Similarly, evaluate $\left(\Delta x^{(j)}\right)^T H \Delta x^{(j)}$.

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):

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

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 expansive operation is therefore computing the gradient which takes time $O(n^2)$.

2 Quasi-Newton Methods

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

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

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:

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

where $p^{(k)} = x^{(k+1)} - x^{(k)}$ and $q^{(k)} = \nabla f\left(x^{(k+1)}\right) - \nabla f\left(x^{(k)}\right)$. 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:

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

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:

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

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

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

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

1. Start from some $x^{(0)}$ and an initial $D^{(0)}$
2. For $i \in \{0, 1, 2, \ldots\}$
3. $\quad \Delta x^{(i)} \leftarrow -D^{(i)} \nabla f\left(x^{(i)}\right)$
4. $\quad t^{(i)} \leftarrow \arg\min_t f\left(x^{(i)} + t \Delta x^{(i)}\right)$
5. $\quad x^{(i+1)} \leftarrow x^{(i)} + t^{(i)} \Delta x^{(i)}$
6. $\quad$ Calculate $D^{(i+1)}$ according to equation (7)

2.2

We now consider applying BFGS to a quadratic objective $f(x) = \frac{1}{2}x'Hx - b'x$ with $x \in \mathbb{R}^n$ and H 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 $D^{(n)} = H^{-1}$, i.e. after n iterations the correct Hessian is recovered.

2.2.2

Show that $\Delta x^{(0)}, \ldots, \Delta x^{(n-1)}$ are H-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.