Problem Set #6, Algorithms for Unconstrained Nonlinear Optimization

OSM Lab: Math
Rebekah Dix

Exercise 1

Claim 1. An unconstrained linear objective function is either constant or has no minimum.

Proof. Let L be a linear objective function, and suppose that L is not constant. For the sake of contradiction, suppose x^* is a minimizer of L. Since L is not constant, there exists an x such that $Lx \neq Lx^*$. There are two cases to consider. If $Lx < Lx^*$, then we have a contradiction as x^* is not a minimizer. Similarly, if $Lx > Lx^*$, then $L(x - x^*) > 0$. Observe that,

$$L(x^* + x^* - x) = Lx^* - L(x - x^*) < Lx^* \tag{1}$$

Therefore x^* is not a minimizer. Thus, if L is not constant, then it has no minimum. Conversely, if L is constant, then it has a minimum, which is the constant value of the function. □

Exercise 2

Claim 2. Let $b \in \mathbb{R}^m$ and $A \in M_{m\times n}(\mathbb{R})$. Then the problem of finding an $x^* \in \mathbb{R}^n$ to minimize $\|Ax - b\|_2$ is equivalent to minimizing

$$x^T A^T A x - 2b^T A x \tag{2}$$

Furthermore, minimizing Equation 2 is equivalent to solving the normal equation

$$A^T A x = A^T b \tag{3}$$

Proof. First observe that,

$$\begin{aligned}
\|Ax - b\|_2^2 &= \langle Ax - b, Ax - b\rangle \\
&= (Ax - b)^T (Ax - b) \\
&= x^T A^T A x - x^T A b - b^T A x + b^T b \\
&= x^T A^T A x - 2b^T A x + b^T b \qquad \text{(because } x^T Ab \text{ and } b^T Ax \text{ are scalar)}
\end{aligned}$$

Next observe that $b^T b$ is a constant. Therefore, minimizing $\|Ax - b\|_2$ is equivalent to minimizing $x^T A^T A x - 2b^T A x$.

All x are interior, so the FONC requires that at a minimum, $2x^T A^T A - 2b^T A = 0$, or that, after taking the transpose of this condition, $A^T A x = A^T b$. Additionally, since $A^T A$ is positive semidefinite, the second order condition will be satisfied. Therefore, minimizing Equation 2 is equivalent to solving the normal equation. □

Exercise 3

Gradient Descent

- **Basic idea:** The negative of the gradient of a function is the direction of greatest decrease. Therefore, to minimize a function, we choose an initial point and continually move in the direction of greatest decrease.
- **Optimization problems it can and cannot solve:** This method requires the objective function to be differentiable.
- **Relative strengths:** This method converges quickly for problems with a low condition number.
- **Relative weaknesses:** This method converges slowly for problems with a large condition number.

Newton and Quasi-Newton Methods

- **Basic idea:** Newton's method creates a local quadratic approximation to the function using the gradient and Hessian at a specific point, and then iterates forward to the minimum of that quadratic approximation.
- **Optimization problems it can and cannot solve:** Newton's method is particularly powerful for quadratic optimization problems.
- **Relative strengths:** Newton's method converges quickly for problems where the dimension is not too large and the Hessian is positive definite and can be computed easily. If the objective function is a quadratic form with a symmetric and positive definite matrix, then Newton's method reaches the optimizer from any starting point in just one iteration.
- **Relative weaknesses:** Newton's method may not converge when the starting point is far from the minimizer, the Hessian is not positive definite, or the inverse of the Hessian is expensive, unstable, or otherwise difficult to compute. Newton's method is also very computationally intensive when the dimension of the probblem is large.

Conjugate Gradient

- **Basic idea:** The conjugate gradient method moves toward the minimizer of a function by moving along Q-conjugate directions.
- **Optimization problems it can and cannot solve:** Conjugate gradient works well for large quadratic minimization problems with sparse matrices.
- **Relative strengths:** Each step of conjugate gradient is generally less expensive to compute than a step in Newton's method. This method never computes or stores the Hessian or approximations to the Hessian. This method is useful

for solving large quadratic optimization problems where the matrix is symmetric, positive definite, and sparse. With sparsity, many iterations of conjugate gradient can be less expensive than a single iteration of Newton's method.

- **Relative weaknesses:** If the matrices in the objective function are not sparse, then each iteration of conjugate gradient can be very expensive.

Exercise 4

Claim 3. Let $f(x) = \frac{1}{2}x^TQx - b^Tx$, where $Q \in M_n(\mathbb{R})$ satisfies $Q > 0$ and $b \in \mathbb{R}^n$. The Method of Steepest Descent converges in one step (i.e. $x_1 = Q^{-1}b$) if and only if x_0 is chosen such that $Df(x_0)^T = Qx_0 - b$ is an eigenvector of Q (and α is chosen optimally).

Proof. First suppose that the Method of Steepest Descent converges in one step. Then,

$$x_1 = Q^{-1}b = x_0 - \alpha_0(Qx_0 - b) \tag{4}$$

for some α_0. Multiply both sides of the equation by Q. Then,

$$b = Qx_0 - \alpha_0 Q(Qx_0 - b) \tag{5}$$

or

$$Q(Qx_0 - b) = \frac{1}{\alpha_0}(Qx_0 - b) \tag{6}$$

Therefore, $(Qx_0 - b)$ is an eigenvector of Q. Clearly, since the method converged in one step, α was chosen optimally and must satisfy (9.2). Conversely, suppose that $Df(x_0)^T = Qx_0 - b$ is an eigenvector of Q with eigenvalue λ and α is chosen optimally according to 9.2. The Method of Steepest Descent implies that,

$$\begin{aligned}
x_1 &= x_0 - \alpha_0 Df(x_0)^T \\
&= x_0 - \frac{Df(x_0)Df(x_0)^T}{Df(x_0)QDf(x_0)^T}Df(x_0)^T \\
&= x_0 - \frac{Df(x_0)Df(x_0)^T}{\lambda Df(x_0)Df(x_0)^T}Df(x_0)^T \\
&= x_0 - \frac{1}{\lambda}Df(x_0)^T \\
&= x_0 - \frac{1}{\lambda}(Qx_0 - b)
\end{aligned}$$

Now, multiply both sides by Q:

$$\begin{aligned}
Qx_1 &= Qx_0 - \frac{1}{\lambda}Q(Qx_0 - b) \\
&= Qx_0 - \frac{1}{\lambda}\lambda(Qx_0 - b) \\
&= Qx_0 - (Qx_0 - b) \\
&= b
\end{aligned}$$

Thus, $x_1 = Q^{-1}b$ so that the Method of Steepest Descent converges in one step. □

Exercise 5

Claim 4. Let $\{x_k\}_{k=0}^{\infty}$ be defined by the Method of Steepest Descent. Then, $x_{k+1}-x_k$ is orthogonal to $x_{k+2}-x_{k+1}$ for each k.

Proof. Recall that in the Method of Steepest Descent, α_k is chosen to minimize $f(x_k-\alpha_k Df(x_k)^T)$. Therefore, by the FONC at the optimal α_k, we have that,

$$Df(x_k-\alpha_k Df(x_k)^T)Df(x_k)^T=0 \tag{7}$$

Also recall that $x_{k+1}-x_k=-\alpha_k Df(x_k)^T$ and $x_{k+2}-x_{k+1}=-\alpha_{k+1}Df(x_{k+1})^T$. Thus,

$$\begin{aligned}\langle x_{k+1}-x_k, x_{k+2}-x_{k+1}\rangle &= \alpha_k\alpha_{k+1}Df(x_k)Df(x_{k+1})^T\\ &= \alpha_k\alpha_{k+1}Df(x_k-\alpha_k Df(x_k)^T)Df(x_k)^T\\ &= 0\end{aligned}$$

Therefore, $x_{k+1}-x_k$ is orthogonal to $x_{k+2}-x_{k+1}$ for each k. □

Exercises 6 - 9: See Jupyter Notebook

Exercise 10

Claim 5. Let $f(x)=\frac{1}{2}x^TQx-b^Tx$ where $Q\in M_n(\mathbb{R})$ is symmetric and positive definite. For any initial guess $x_0\in R^n$, one iteration of Newton's method lands at the unique minimizer of f.

Proof. First observe that $Df(x_0)^T=Qx_0-b$ and $D^2f(x_0)=Q$. Therefore,

$$\begin{aligned}x_1 &= x_0-D^2f(x_0)^{-1}Df(x_0)^T\\ &= x_0-Q^{-1}(Qx_0-b)\\ &= x_0-x_0+Q^{-1}b\\ &= Q^{-1}b\end{aligned}$$

Then, since Q is positive definite, this system of equations will have a unique solution so that $x_1=Q^{-1}b$ is the unique minimizer of f by the FONC and second-order sufficient condition. □

Exercise 12

Claim 6. If $A\in M_n(\mathbb{F})$ has eigenvalues $\lambda_1,\ldots,\lambda_n$ and $B=A+\mu I$, then the eigenvectors of A and B are the same, and the eigenvalues of B are $\mu+\lambda_1,\mu+\lambda_2,\ldots,\mu+\lambda_n$.

Proof. Let λ_i and x_i be an eigenvalue and eigenvector, respectively, of A. Then,

$$\begin{aligned}Bx_i &= (A+\mu I)x_i\\ &= Ax_i+\mu Ix_i\\ &= \lambda_i x_i+\mu x_i\\ &= (\lambda_i+\mu)x_i\end{aligned}$$

Therefore, $(\lambda_i + \mu)$ is an eigenvalue of B with eigenvector x_i. It follows that A and B have the same eigenvectors, and the eigenvalues of B are $\mu+\lambda_1, \mu+\lambda_2, \ldots, \mu+\lambda_n$. $\square$

Exercise 15

Claim 7 (Sherman-Morrison-Woodbury)**.** Let A be a nonsingular $n \times n$ matrix, B an $n \times l$ matrix, C a nonsingular $l \times l$ matrix, and D an $l \times n$ matrix. We have,

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \tag{8}$$

Proof. To prove this identity, we show that $(A+BCD)(A^{-1}-A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}) = I$. To that end,

$$\begin{aligned}
&(A + BCD)(A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}) \\
&= AA^{-1} - AA^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \\
&\quad + BCDA^{-1} - BCDA^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \\
&= I + BCDA^{-1} \\
&\quad - (B(C^{-1} + DA^{-1}B)^{-1} + BCDA^{-1}B(C^{-1} + DA^{-1}B)^{-1})DA^{-1} \\
&= I + BCDA^{-1} \\
&\quad - (BCC^{-1} + BCDA^{-1}B)(C^{-1} + DA^{-1}B)^{-1})DA^{-1} \\
&= I + BCDA^{-1} \\
&\quad - BC(C^{-1} + DA^{-1}B)(C^{-1} + DA^{-1}B)^{-1})DA^{-1} \\
&= I + BCDA^{-1} - BCDA^{-1} \\
&= I
\end{aligned}$$

Therefore, $(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$. $\square$

Exercise 16

We have that

$$A_{k+1} = A_k + \frac{y_k - A_k s_k}{\|s_k\|^2} s_k^T \tag{9}$$

We can use the Sherman-Morrison-Woodbury Identity to invert A_{k+1}. First observe that A_k is $n \times n$, $y_k - A_k s_k$ is $n \times 1$, $\|s_k\|^2$ is 1×1, and s_k^T is $1 \times n$. These dimensions suggest the following correspondence for applying the identity: $A = A_k$, $B = y_k - A_k s_k$, $C = \frac{1}{\|s_k\|^2}$, and $D = s_k^T$. Then,

$$\begin{aligned}
A_{k+1}^{-1} &= (A + BCD)^{-1} \\
&= A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \\
&= A_k^{-1} - A_k^{-1}\frac{(y_k - A_k s_k)s_k^T A_k^{-1}}{(\|s_k\|^2 + s_k^T A_k^{-1}(y_k - A_k s_k))} \\
&= A_k^{-1} + \frac{(s_k - A_k^{-1}y_k)s_k^T A_k^{-1}}{s_k^T A_k^{-1} y_k}
\end{aligned}$$

Exercise 18

Claim 8. Let $Q \in M_n(\mathbb{R})$ satisfy $Q > 0$, and let f be the quadratic function $f(x) = \frac{1}{2}x^TQx - b^Tx + c$. Given a starting point x_0 and Q-conjugate directions $d_0, d_1, \ldots, d_{n-1}$ in $\mathbb{R}^n$, the optimal line search solution for $x_{k+1} = x_k + \alpha_k d_k$ is given by $\alpha_k = \frac{r_k^T d_k}{d_k^T Q d_k}$, where $r_k = b - Qx_k$.

Proof. The optimal line search solution for $x_{k+1} = x_k + \alpha_k d_k$ must satisfy $0 = f'(x_k + \alpha_k d_k)$. Then,

$$\begin{aligned} 0 &= f'(x_k + \alpha_k d_k) \\ &= Df(x_k + \alpha_k d_k)d_k \\ &= ((x_k + \alpha_k d_k)^T Q - b^T)d_k \\ &= x_k^T Q d_k + \alpha_k d_k^T Q d_k - b^T d_k \end{aligned}$$

Solving this equation for α_k yields, $\alpha_k = \frac{r_k^T d_k}{d_k^T Q d_k}$, where $r_k = b - Qx_k$. □

Exercise 20

Claim 9. In the Conjugate Gradient Algorithm, $r_i^T r_k = 0$ for all $i < k$.

Proof. I prove this statement by induction. Recall that $r_0 = d_0$. Then,

$$\begin{aligned} r_0^T r_1 &= r_0^T(b - Qx_1) \\ &= r_0^T(b - Qx_0 - a_0 Q d_0) \\ &= r_0^T r_0 - a_0 r_0^T A r_0 \\ &= r_0^T r_0 - \frac{r_0^T r_0}{r_0^T Q r_0} r_0^T A r_0 \\ &= r_0^T r_0 - r_0^T r_0 = 0 \end{aligned}$$

Therefore, $r_0^T r_1$. This pattern will continue for all $i < k$. Thus, by induction, we have that $r_i^T r_k = 0$ for all $i < k$. □