OSM 2018 MATH PROBLEM SET 6 DANIEL FOX

9.1

Proof. Let f be an unconstrained linear objective function. Suppose by way of contradiction that $\mathbf{x}^*$ is a minimizer of f. Given that f is not constant, it must be that $f\mathbf{x}^* \neq f\mathbf{x}$ for some $\mathbf{x}$.
Now if $f\mathbf{x}^* > f\mathbf{x}$, it follows by contradiction that $\mathbf{x}^*$ is not a minimizer.
Now if $f\mathbf{x}^* < f\mathbf{x}$, it follows that $f(\mathbf{x} - \mathbf{x}^*) > 0$, and hence

$$\begin{aligned} f(\mathbf{x}^* + \mathbf{x}^* - \mathbf{x}) &= f\mathbf{x}^* - f(\mathbf{x} - \mathbf{x}^*) \\ &\leq f\mathbf{x}^* \end{aligned}$$

and $\mathbf{x}^*$ is not a minimizer. Thus f has no minimizer. By the converse, if f is constant, it must have a minimum. □

9.2

Proof. Let $\mathbf{b} \in \mathbb{R}^m$ and $A \in M_{m\times n}(\mathbb{R})$. Observe that minimizing $||A\mathbf{x} - \mathbf{b}||_2$ is equivalent to minimizing

$$\begin{aligned} (A\mathbf{x} - \mathbf{b})^T(A\mathbf{x} - \mathbf{b}) &= (\mathbf{x}^T A^T - \mathbf{b}^T)(A\mathbf{x} - \mathbf{b}) \\ &= \mathbf{x}^T A^T A\mathbf{x} - \mathbf{x}^T A^T\mathbf{b} - \mathbf{b}^T A\mathbf{x} + 2\mathbf{b}^T\mathbf{b} \\ &= \mathbf{x}^T A^T A\mathbf{x} - 2\mathbf{b}^T A\mathbf{x} + 2\mathbf{b}^T\mathbf{b}, \quad \text{where } A^T A \geq 0 \end{aligned}$$

By the first order condition, we have that

$$\begin{aligned} 2\mathbf{x}^T A^T A - 2\mathbf{b}^T A &= 0 \\ 2\mathbf{x}^T A^T A &= 2\mathbf{b}^T A \\ \mathbf{x}^T A^T A &= \mathbf{b}^T A \\ A^T A\mathbf{x} &= A^T\mathbf{b} \end{aligned}$$

By the second order condition, we have that $2A^T A > 0$ since $A^T A \geq 0$. □

9.3

(i) Gradient: We pick some initial point and move toward the direction of the greatest decrease (the negative of the gradient of a function).

Newton's: Uses local quadratic approximation with the gradient and Hessian, and then iterately solves the problem.

BFGS: Begins with solving for the Hessian, but then uses an approximation of it to save on computational time. Assumes the gradient of the approximation is close to the approximation of gradient of f at $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$.

Conjugate Gradient: Finds the minimizer by moving along Q-conjugate directions.

(ii) Gradient: Requires the objective function to be differentiable.

Newton's: Good for quadratic optimization problems, as long as the dimensionality is not too large.

BFGS: Can be useful for non-smooth optimizations.

Conjugate Gradient: Works well for large quadratic minimization problems where matrices are sparse.

(iii) Gradient: Less expensive per iteration.

Newton's: Converges quickly.

BFGS: Less expensive per iteration than Newton's method.

Conjugate Gradient: Generally less expensive than Newton's method. Good for solving relatively larger quadratic optimization problems.

(iv) Gradient: Slow convergence

Newton's: Each iteration is expensive

BFGS: Expensive to store the approximations

Conjugate Gradient: Can be expensive if the matrices of the objective function are not sparse.

9.4

Proof. Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^TQ\mathbf{x} - \mathbf{b}^T\mathbf{x}$ where $Q \in M_n(\mathbb{R})$ satisfies $Q > 0$ and $\mathbf{b} \in \mathbb{R}^n$. Now suppose the Method of Steepest Descent converges in one step.

That is $\mathbf{x}_1 = Q^{-1}\mathbf{b}f$. Note that

$$\begin{aligned}\mathbf{x}_1 &= Q^{-1}\mathbf{b}\\ &= \mathbf{x}_0 - \alpha_0(Q\mathbf{x}_0 - \mathbf{b})\end{aligned}$$

given some α_0. So,

$$\mathbf{b} = Q\mathbf{x}_0 - \alpha_0 Q(Q\mathbf{x}_0 - \mathbf{b})$$

Rearranging terms gives us

$$Q(Q\mathbf{x}_0 - \mathbf{b}) = \frac{1}{\alpha_0}(Q\mathbf{x}_0 - \mathbf{b})$$

where $Q\mathbf{x}_0 - \mathbf{b}$ is an eigenvector of Q. Thus, α_0 must have been chosen to satisfy (9.2).
Now let $Df(\mathbf{x}_0)^T = Q\mathbf{x}_0 - \mathbf{b}$ be an eigenvector of Q where α_0 satisfies (9.2). Given an eigenvalue λ, by the Method of Steepest Descent

$$\begin{aligned}\mathbf{x}_1 &= \mathbf{x}_0 - \alpha_0 Df(\mathbf{x}_0)^T\\ &= \mathbf{x}_0 - \frac{Df(\mathbf{x}_0)Df(\mathbf{x}_0)^T}{Df(\mathbf{x}_0)QDf(\mathbf{x}_0)^T}Df(\mathbf{x}_0)^T\\ &= \mathbf{x}_0 - \frac{Df(\mathbf{x}_0)Df(\mathbf{x}_0)^T}{\lambda Df(\mathbf{x}_0)Df(\mathbf{x}_0)^T}Df(\mathbf{x}_0)^T\\ &= \mathbf{x}_0 - \frac{1}{\lambda}Df(\mathbf{x}_0)^T\\ &= \mathbf{x}_0 - \frac{1}{\lambda}(Q\mathbf{x}_0 - \mathbf{b})\\ Q\mathbf{x}_1 &= Q\mathbf{x}_0 - \frac{1}{\lambda}Q(Q\mathbf{x}_0 - \mathbf{b})\\ &= Q\mathbf{x}_0 - (Q\mathbf{x}_0 - \mathbf{b})\\ &= \mathbf{b}\\ \mathbf{x}_1 &= Q^{-1}\mathbf{b}\end{aligned}$$

and the Method of Steepest Descent converges in one step. □

9.5

Proof. Let $f : \mathbb{R}^n \to \mathbb{R}$ is C^1. Let $\{\mathbf{x}_k\}_{k=0}^{\infty}$ be defined by the Method of Steepest Descent. Given an optimal choice of α_k, we have by the first order necessary condition that

$$Df(\mathbf{x}_k - \alpha_k Df(\mathbf{x}_k)^T) Df(\mathbf{x}_k)^T = 0$$

Furthermore, we know that $\mathbf{x}_{k+1} - \mathbf{x}_k = -\alpha_k Df(\mathbf{x}_k)^T$ and $\mathbf{x}_{k+2} - \mathbf{x}_{k+1} = -\alpha_{k+1} Df(\mathbf{x}_{k+1})^T$. Observe that the inner product between these two differences yields

$$\begin{aligned} \alpha_k \alpha_{k+1} Df(\mathbf{x}_k) Df(\mathbf{x}_{k+1})^T &= \alpha_k \alpha_{k+1} Df(\mathbf{x}_k - \alpha_k Df(\mathbf{x}_k)^T) Df(\mathbf{x}_k)^T \\ &= 0 \end{aligned}$$

and $\mathbf{x}_{k+1} - \mathbf{x}_k$ is orthogonal to $\mathbf{x}_{k+2} - \mathbf{x}_{k+1}$ for each k. □

9.6 See Jupyter notebook.
9.7 See Jupyter notebook.
9.8 See Jupyter notebook.
9.9 See Jupyter notebook.
9.10 Consider the quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q \mathbf{x} - \mathbf{b}^T \mathbf{x}$, where $Q \in M_n(\mathbb{R})$ is symmetric and positive definition and $\mathbf{b} \in \mathbb{R}^n$. Let $\mathbf{x}_0 \in \mathbb{R}^n$ be an initial guess. Observe that

$$\begin{aligned} Df &= Q\mathbf{x} - \mathbf{b} \\ D^2 f &= Q \end{aligned}$$

Thus, a minimum is achieved whenever $Df = Q\mathbf{x} - \mathbf{b} = 0$. Now, applying

Newton's method, we see that

$$\begin{aligned}
\mathbf{x}_1 &= \mathbf{x}_0 - Q^{-1}Df(\mathbf{x}_0) \\
&= \mathbf{x}_0 - Q^{-1}(Q\mathbf{x}_0 - \mathbf{b}) \\
&= \mathbf{x}_0 - Q^{-1}Q\mathbf{x}_0 + Q^{-1}\mathbf{b} \\
&= \mathbf{x}_0 - \mathbf{x}_0 + Q^{-1}\mathbf{b} \\
&= Q^{-1}\mathbf{b} \\
Q\mathbf{x}_1 &= \mathbf{b} \\
Q\mathbf{x}_1 - \mathbf{b} &= 0
\end{aligned}$$

So, $\mathbf{x}_1$ is the minimum; therefore, one iteration of Newton's method lands at the unique minimizer of f.

9.12

Proof. Let $A \in M_n(\mathbb{F})$ with eigenvalues $\lambda_1, ..., \lambda_n$, and $B = A + \mu I$. Observe that, for any given eigenvalue λ_i and corresponding eigenvector v_i we have

$$\begin{aligned}
Bv_i &= (A + \mu I)v_i \\
&= Av_i + \mu I v_i \\
&= \lambda_i v_i + \mu v_i \\
&= (\lambda_i + \mu)v_i
\end{aligned}$$

Therefore, the eigenvectors of A and B are the same, and the eigenvalues of B are $\mu + \lambda_1, \mu + \lambda_2, ..., \mu + \lambda_n$ □

9.15

Proof. Let A be a nonsingular $n \times n$ matrix, B an n matrix, C a nonsingular

$\ell \times \ell$ matrix, and D an $\ell \times n$ matrix. Observe that

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

Since $(A+BCD)(A^{-1}-A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}) = I$, it follows that

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

□

9.16

From the Quasi-Newton method, we have that

$$A_{k+1} = A_k + \frac{\mathbf{y}_k - A_k\mathbf{s}_k}{||\mathbf{s}_k||^2}\mathbf{s}_k^T$$

Using (9.13) which is the Sherman-Morrison-Woodbury formula, we can let $A = A_k, B = \mathbf{y}_k - A_k\mathbf{s}_k, C = \frac{1}{||\mathbf{s}_k||^2}, \text{and} D = \mathbf{s}_k^T$. observe that

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

as desired.

9.18

Proof. Let $Q \in M_n(\mathbb{R})$ satisfy $Q > 0$, and let f be the quadratic function $f(x) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} - \mathbf{b}^T\mathbf{x} + c$. Now, let α_k minimize $\phi_k(\alpha) = f(\mathbf{x}_k + \alpha_k\mathbf{d}_k)$. Observe that

$$\begin{aligned} 0 &= \phi_k'(\alpha_k) \\ &= -Df(\mathbf{x}_k + \alpha_k\mathbf{d}_k)\mathbf{d}_k \\ &= ((\mathbf{x}_k - \alpha_k\mathbf{d}_k)^T Q - \mathbf{b}^T)\mathbf{d}_k \\ &= (\mathbf{x}_k^T Q - \mathbf{b}^T)\mathbf{d}_k - (\alpha_k\mathbf{d}_k)^T Q\mathbf{d}_k \\ &= \mathbf{r}_k^T\mathbf{d}_k - \alpha_k(\mathbf{d}_k^T Q\mathbf{d}_k) \end{aligned}$$

where $\mathbf{r}_k = \mathbf{b} - Q\mathbf{x}_k$. Now, since $0 = \mathbf{r}_k^T\mathbf{d}_k - \alpha_k(\mathbf{d}_k^T Q\mathbf{d}_k)$, it follows that

$$\begin{aligned} \alpha_k(\mathbf{d}_k^T Q\mathbf{d}_k) &= \mathbf{r}_k^T\mathbf{d}_k \\ \alpha_k &= \frac{\mathbf{r}_k^T\mathbf{d}_k}{\mathbf{d}_k^T Q\mathbf{d}_k} \end{aligned}$$

as desired. □

9.20

Proof. By the Conjugate Gradient Algorithm, initialize by setting $k = 0$, and select the initial point $\mathbf{x}_0$. Set $\mathbf{d}_0 = \mathbf{r}_0 = -Df(\mathbf{x}_0)^T$. Part (i) of the algorithm states that

$$\alpha_k = \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r_k}^T Q\mathbf{r}_k}$$

and part (ii) states that

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{r}_k$$

while part (iii) states that

$$\mathbf{r}_{k+1} = \mathbf{b} - Q\mathbf{x}_{k+1}$$

Now let $k = 0$. Using these facts, observe that

$$\begin{aligned}
\alpha_0 &= \frac{\mathbf{r}_0^T \mathbf{r}_0}{\mathbf{r_0}^T Q \mathbf{r}_0} \\
\mathbf{r}_0^T \mathbf{r}_0 &= \alpha_0 \mathbf{r_0}^T (Q \mathbf{r}_0) \\
(\mathbf{b} - Q\mathbf{x}_0)^T \mathbf{r}_0 &= \alpha_0 (Q\mathbf{r}_0)^T \mathbf{r}_0 \\
(\mathbf{b} - Q\mathbf{x}_0)^T \mathbf{r}_0 - \alpha_0 (Q\mathbf{r}_0)^T \mathbf{r}_0 &= 0 \\
(\mathbf{b} - Q(\mathbf{x}_0 + \alpha_0 \mathbf{r}_0))^T \mathbf{r}_0 &= 0 \\
(\mathbf{b} - Q\mathbf{x}_1)^T \mathbf{r}_0 &= 0 \\
\mathbf{r}_1^T \mathbf{r}_0 &= 0
\end{aligned}$$

Taking the transpose of both sides yields

$$\mathbf{r}_0^T \mathbf{r}_1 = 0$$

By induction, we can continue to increment. Therefore $\mathbf{r}_i^T \mathbf{r}_k = 0$ for all $i < k$ □