# EE/M 520, Spring 2007
## Exam 1: Session 14

**Solutions** (version: March 6, 2007, 9:6) 75 mins.; Total 50 pts.

**1.** (15 pts.) Consider the set-constrained problem

$$\begin{array}{ll} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \Omega, \end{array}$$

where $\Omega = \{[x_1, x_2]^T : x_1^2 + x_2^2 = 1\}$.

a. Consider a point $\boldsymbol{x}^* \in \Omega$. Specify all feasible directions at $\boldsymbol{x}^*$.

b. Which points in $\Omega$ satisfy the FONC for this set-constrained problem?

c. Based on part b, is the FONC for this set-constrained problem useful for eliminating local-minimizer candidates?

d. Suppose we use polar coordinates to parameterize points $\boldsymbol{x} \in \Omega$ in terms of a single parameter $\theta$:

$$x_1 = \cos\theta \qquad x_2 = \sin\theta.$$

Now use the FONC for unconstrained problems (with respect to $\theta$) to derive a necessary condition of this sort: if $\boldsymbol{x}^* \in \Omega$ is a local minimizer, then $\boldsymbol{d}^T \nabla f(\boldsymbol{x}^*) = 0$ for all $\boldsymbol{d}$ satisfying a "certain condition." Specify what this "certain condition" is.

**Ans.:** a. There are no feasible directions at any $\boldsymbol{x}^*$.

b. *All* points in $\Omega$ satisfy the FONC for this set-constrained problem.

c. No, the FONC for this set-constrained problem is not useful for eliminating local-minimizer candidates.

d. Write $h(\theta) = f(g(\theta))$, where $g : \mathbb{R} \to \mathbb{R}^2$ is given by the equations relating $\theta$ to $\boldsymbol{x} = [x_1, x_2]^T$. Note that $Dg(\theta) = [-\sin\theta, \cos\theta]^T$. Hence, by the chain rule,

$$h'(\theta) = Df(g(\theta))Dg(\theta) = Dg(\theta)^T \nabla f(g(\theta)).$$

Notice that $Dg(\theta)$ is tangent to $\Omega$ at $\boldsymbol{x} = g(\theta)$. Alternatively, we could say that $Dg(\theta)$ is orthogonal to $\boldsymbol{x} = g(\theta)$.

Suppose $\boldsymbol{x}^* \in \Omega$ is a local minimizer. Write $\boldsymbol{x}^* = g(\theta^*)$. Then $\theta^*$ is an unconstrained minimizer of $h$. By the FONC for unconstrained problems, $h'(\theta^*) = 0$, which implies that $\boldsymbol{d}^T \nabla f(\boldsymbol{x}^*) = 0$ for all $\boldsymbol{d}$ tangent to $\Omega$ at $\boldsymbol{x}^*$ (or, alternatively, for all $\boldsymbol{d}$ orthogonal to $\boldsymbol{x}^*$).

**2.** (12 pts.) Suppose we wish to solve the equation $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$, where

$$\boldsymbol{h}(\boldsymbol{x}) = \begin{bmatrix} 4 + 3x_1 + 2x_2 \\ 1 + 2x_1 + 3x_2 \end{bmatrix}.$$

Consider using an algorithm of the form $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha \boldsymbol{h}(\boldsymbol{x}^{(k)})$, where $\alpha$ is scalar constant that does not depend on $k$.

a. Find the solution of $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$.

b. Find the largest range of values of $\alpha$ such that the algorithm is globally convergent to the solution of $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$.

c. Assuming that $\alpha$ is outside the range of values in part b, give an example of an initial condition $\boldsymbol{x}^{(0)}$ of the form $[x_1, 0]^T$ such that the algorithm is guaranteed not satisfy the descent property.

**Ans.:** a. We can write $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}$, where $\boldsymbol{b} = [-4, -1]^T$ and

$$\boldsymbol{Q} = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$$

is positive definite. Hence, the solution is

$$\boldsymbol{Q}^{-1}\boldsymbol{b} = \frac{1}{5}\begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix}\begin{bmatrix} -4 \\ -1 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}.$$

b. By part a, the algorithm is a fixed-step-size gradient algorithm for a problem with gradient $\boldsymbol{h}$. The eigenvalues of $\boldsymbol{Q}$ are $1$ and $5$. Hence, the largest range of values of $\alpha$ such that the algorithm is globally convergent to the solution is $0 < \alpha < 2/5$.

c. The eigenvectors of $\boldsymbol{Q}$ corresponding to eigenvalue $5$ has the form $c[1, 1]^T$, where $c \in \mathbb{R}$. Hence, to violate the descent property, we pick

$$\boldsymbol{x}^{(0)} = \boldsymbol{Q}^{-1}\boldsymbol{b} + c\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -3 \\ 0 \end{bmatrix}$$

where we choose $c = -1$ so that $\boldsymbol{x}^{(0)}$ has the specified form.

**3.** (7 pts.) Consider a scalar sequence $\{x_k\}$ that converges with order of convergence $p$, and satisfies

$$\lim_{k \to \infty} \frac{|x_{k+1} - 2|}{|x_k - 2|^3} = 0.$$

What is the limit of $\{x_k\}$? What can you say about $p$?

**Ans.:** The limit of $\{x_k\}$ must be $2$, because it is clear from the given equation that $|x_{k+1} - 2| \to 0$. Also, we see that $|x_{k+1} - 2| = o(|x_k - 2|^3)$. Hence, we conclude that $p > 3$.

**4.** (16 pts.) Let $f : \mathbb{R}^n \to \mathbb{R}$ be such that $f \in \mathcal{C}^1$. Consider an optimization algorithm applied to this $f$, of the usual form $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}$, where $\alpha_k \geq 0$ is chosen according to line search. Suppose that $\boldsymbol{d}^{(k)} = -\boldsymbol{H}_k \boldsymbol{g}^{(k)}$, where $\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)})$ and $\boldsymbol{H}_k$ is symmetric.

a. Show that if $\boldsymbol{H}_k$ satisfies the following conditions whenever the algorithm is applied to a quadratic, then the algorithm is quasi-Newton:

i. $\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \boldsymbol{U}_k$

ii. $\boldsymbol{U}_k \Delta \boldsymbol{g}^{(k)} = \Delta \boldsymbol{x}^{(k)} - \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)}$

iii. $\boldsymbol{U}_k = \boldsymbol{a}^{(k)} \Delta \boldsymbol{x}^{(k)T} + \boldsymbol{b}^{(k)} \Delta \boldsymbol{g}^{(k)T} \boldsymbol{H}_k$, where $\boldsymbol{a}^{(k)}$ and $\boldsymbol{b}^{(k)}$ are in $\mathbb{R}^n$

b. Which (if any) among the rank-one, DFP, and BFGS algorithms satisfy the three conditions in part a (whenever the algorithm is applied to a quadratic)? (For those that do, you must specify the vectors $\boldsymbol{a}^{(k)}$ and $\boldsymbol{b}^{(k)}$.)

**Ans.:** a. Suppose the three conditions hold whenever applied to a quadratic. We need to show that when applied to a quadratic, for $k = 0, \dots, n-1$ and $i = 0, \dots, k$, $\boldsymbol{H}_{k+1} \Delta \boldsymbol{g}^{(i)} = \Delta \boldsymbol{x}^{(i)}$. For $i = k$, we have

$$\begin{aligned} \boldsymbol{H}_{k+1} \Delta \boldsymbol{g}^{(k)} &= \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)} + \boldsymbol{U}_k \Delta \boldsymbol{g}^{(k)} \quad \text{by condition i} \\ &= \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)} + \Delta \boldsymbol{x}^{(k)} - \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)} \quad \text{by condition ii} \\ &= \Delta \boldsymbol{x}^{(k)}, \end{aligned}$$

as required. For the rest of the proof ($i = 0, \dots, k-1$), we use induction on $k$.

For $k = 0$, there is nothing to prove (covered by the $i = k$ case). So suppose the result holds for $k-1$. To show the result for $k$, first fix $i \in \{0, \dots, k-1\}$. We have

$$\begin{aligned} \boldsymbol{H}_{k+1} \Delta \boldsymbol{g}^{(i)} &= \boldsymbol{H}_k \Delta \boldsymbol{g}^{(i)} + \boldsymbol{U}_k \Delta \boldsymbol{g}^{(i)} \\ &= \Delta \boldsymbol{x}^{(i)} + \boldsymbol{U}_k \Delta \boldsymbol{g}^{(i)} \quad \text{by the induction hypothesis} \\ &= \Delta \boldsymbol{x}^{(i)} + \boldsymbol{a}^{(k)} \Delta \boldsymbol{x}^{(k)T} \Delta \boldsymbol{g}^{(i)} + \boldsymbol{b}^{(k)} \Delta \boldsymbol{g}^{(k)T} \boldsymbol{H}_k \Delta \boldsymbol{g}^{(i)} \quad \text{by condition iii.} \end{aligned}$$

So it suffices to show that the second and third terms are both $0$. For the second term,

$$\begin{aligned} \Delta \boldsymbol{x}^{(k)T} \Delta \boldsymbol{g}^{(i)} &= \Delta \boldsymbol{x}^{(k)T} \boldsymbol{Q} \Delta \boldsymbol{x}^{(i)} \\ &= \alpha_k \alpha_i \boldsymbol{d}^{(k)T} \boldsymbol{Q} \boldsymbol{d}^{(i)} \\ &= 0 \end{aligned}$$

because of the induction hypothesis, which implies $\boldsymbol{Q}$-conjugacy (where $\boldsymbol{Q}$ is the Hessian of the given quadratic). Similarly, for the third term,

$$\begin{aligned} \Delta \boldsymbol{g}^{(k)T} \boldsymbol{H}_k \Delta \boldsymbol{g}^{(i)} &= \Delta \boldsymbol{g}^{(k)T} \Delta \boldsymbol{x}^{(i)} \quad \text{by the induction hypothesis} \\ &= \Delta \boldsymbol{x}^{(k)T} \boldsymbol{Q} \Delta \boldsymbol{x}^{(i)} \\ &= \alpha_k \alpha_i \boldsymbol{d}^{(k)T} \boldsymbol{Q} \boldsymbol{d}^{(i)} \\ &= 0, \end{aligned}$$

again because of the induction hypothesis, which implies $\boldsymbol{Q}$-conjugacy. This completes the proof.

b. All three algorithms satisfy the conditions in part a. Condition i holds, as described in class. Condition ii is straightforward to check for all three algorithms. For the rank-one and DFP algorithms, this is shown in the book. For BFGS, some simple matrix algebra establishes that it holds. Condition iii holds by appropriate definition of the vectors $\boldsymbol{a}^{(k)}$ and $\boldsymbol{b}^{(k)}$. In particular, for the rank-one algorithm,

$$\boldsymbol{a}^{(k)} = \frac{(\Delta \boldsymbol{x}^{(k)} - \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)})}{(\Delta \boldsymbol{x}^{(k)} - \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)})^T \Delta \boldsymbol{g}^{(k)}}, \qquad \boldsymbol{b}^{(k)} = -\frac{(\Delta \boldsymbol{x}^{(k)} - \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)})}{(\Delta \boldsymbol{x}^{(k)} - \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)})^T \Delta \boldsymbol{g}^{(k)}}.$$

For the DFP algorithm,

$$\boldsymbol{a}^{(k)} = \frac{\Delta \boldsymbol{x}^{(k)}}{\Delta \boldsymbol{x}^{(k)T} \Delta \boldsymbol{g}^{(k)}}, \qquad \boldsymbol{b}^{(k)} = -\frac{\boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)}}{\Delta \boldsymbol{g}^{(k)T} \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)}}.$$

Finally, for the BFGS algorithm,

$$\boldsymbol{a}^{(k)} = \left(1 + \frac{\Delta \boldsymbol{g}^{(k)T} \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)}}{\Delta \boldsymbol{g}^{(k)T} \Delta \boldsymbol{x}^{(k)}}\right) \frac{\Delta \boldsymbol{x}^{(k)}}{\Delta \boldsymbol{x}^{(k)T} \Delta \boldsymbol{g}^{(k)}} - \frac{\boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)}}{\Delta \boldsymbol{g}^{(k)T} \Delta \boldsymbol{x}^{(k)}}, \qquad \boldsymbol{b}^{(k)} = \frac{\Delta \boldsymbol{x}^{(k)}}{\Delta \boldsymbol{g}^{(k)T} \Delta \boldsymbol{x}^{(k)}}.$$