1(a). Given an estimate x_k of a local minimizer of $f(x)$, a *linesearch* method (i) computes a search direction s_k, which must also be a descent direction (i.e., $s_k^T \nabla_x f(x_k) < 0$), [1 mark], and (ii) computes a stepsize α_k so that $f(x_k + \alpha_k s_k)$ is "sufficiently" smaller than $f(x_k)$ (using for instance a backtracking Armijo rule). [1 mark]. The next iterate is $x_{k+1} = x_k + \alpha_k s_k$. [1 mark].

The Armijo condition is that the stepsize α_k must satisfy

$$f(x_k + \alpha_k p_k) \leq f(x_k) + \beta \alpha_k p_k^T \nabla_x f(x_k)$$

for some $\beta \in (0, 1)$. [2 marks]. It is important as it stops the stepsize from becoming too long relative to the expected decrease in f. [1 mark].

Let $\mathcal{N} = \{0, 1, 2, \ldots\}$. Given an initial "guess" at the stepsize α_{init} and a sequence of decreasing stepsizes $\{\alpha_{init}\tau^i\}_{i \in \mathcal{N}}$ for some $\tau \in (0, 1)$, a backtracking-Armijo linesearch sets $\alpha_k = \alpha_{init}\tau^l$ where l is the smallest member $\mathcal{N}$ for which

$$f(x_k + \alpha^{(l)} p_k) \leq f(x_k) + \beta \alpha^{(l)} p_k^T \nabla_x f(x_k). \qquad \text{[2 marks]}$$

1(b). The steepest-descent direction is the vector $p_k = -\nabla_x f(x_k)$. [1 mark]. The Newton direction is a solution p_k to the system

$$\nabla_{xx} f(x_k) p_k = -\nabla_x f(x_k)$$

if a solution exists. [1 mark].

Advantages of steepest descent direction [1 mark for at least one of]:

(i) it is cheap (only requires first derivatives),

(ii) it is the archetypical globally convergent method, and

(iii) many other methods resort to steepest descent in bad cases.

Disadvantages of steepest descent direction [1 mark for at least one of]:

(i) it is not scale invariant,

(ii) convergence is usually very (very!) slow (linear), and

(iii) numerically it is often not convergent at all.

Advantages of the Newton direction [1 mark for at least one of]:

(i) it is scale invariant, and

(ii) the iterates usually converge fast (quadratically)

Disadvantages of the Newton direction [1 mark for at least one of]:

(i) it may fail if the Hessian is singular,

(ii) it may give an ascent direction if the Hessian is indefinite

(iii) it is expensive (requires second derivatives and matrix factorization)

1(c). At $x = (\frac{1}{2}, 1)$, the gradient and Hessian are

$$\nabla_x f(x) = \begin{pmatrix} 4x_1^3 - 4x_1 \\ 2x_2 \end{pmatrix} = \begin{pmatrix} -\frac{3}{2} \\ 2 \end{pmatrix}$$

and

$$\nabla_{xx} f(x) = \begin{pmatrix} 12x_1^2 - 4 & 0 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 2 \end{pmatrix} \qquad \text{[1 mark]}$$

Since the product of the Newton direction

$$\begin{pmatrix} -\frac{3}{2} \\ -1 \end{pmatrix}$$

with the gradient is $\frac{1}{4} > 0$, the Newton direction is not a descent direction. [1 mark]

To modify the Newton direction, replace the Newton system by

$$(\nabla_{xx} f(x_k) + M_k) p_k = -\nabla_x f(x_k) \qquad \text{[1 mark]}$$

where M_k is chosen so that $\nabla_{xx} f(x_k) + M_k$ is "sufficiently" positive definite and $M_k = 0$ when $\nabla_{xx} f(x_k)$ is itself "sufficiently" positive definite. [1 mark].

There are various ways of doing this. For example, (i) if $\nabla_{xx} f(x_k)$ has the spectral decomposition $\nabla_{xx} f(x_k) = Q_k D_k Q_k^T$, where Q_k is an orthonormal matrix of eigenvectors, and D_k a diagonal matrix of eigenvalues, then pick

$$\nabla_{xx} f(x_k) + M_k = Q_k \max(\epsilon, |D_k|) Q_k^T$$

for some small $\epsilon > 0$. Alternatively (ii) one could pick $M_k = \max(0, -\lambda_{\min}(\nabla_{xx} f(x_k)))I$ where $\lambda_{\min}(\nabla_{xx} f(x_k)$ is the smallest eigenvalue of $\nabla_{xx} f(x_k)$, or (iii) use a modified Cholesky factorization. [2 marks, for any of these]. So in our case

$$(\nabla_{xx} f(x_k) + M_k) = \begin{pmatrix} \epsilon & 0 \\ 0 & 2 \end{pmatrix}$$

from which the modified Newton direction

$$\begin{pmatrix} -\frac{3}{2\epsilon} \\ -1 \end{pmatrix}$$

is a descent direction [2 marks].

2(a) First-order optimality conditions are that there exist Lagrange multipliers y_* for which x_* is primal feasible, i.e.,

$$c(x_*) \geq 0, \qquad \text{[1 mark]}$$

dual feasible i.e.,

$$\nabla_x f(x_*) - (\nabla c(x_*))^T y_* = 0 \ \text{ and } \ y_* \geq 0, \qquad \text{[1 mark]}$$

and satisfies the complementary slackness condition

$$c_i(x_*)(y_*)_i = 0 \ \text{ for each constraint.} \qquad \text{[1 mark]}$$

2(b). The logarithmic barrier function is

$$\Phi(x, \mu) = f(x) - \mu \sum_{i=1}^{m} \log c_i(x) \qquad \text{[1 mark]}$$

Its gradient is

$$\nabla_x f(x) - (\nabla c(x))^T y(x)$$

where $y_i(x) = \mu / c_i(x)$. [1 mark].

2(c). Let $x(\mu)$ be a local minimizer of the logarithmic barrier function, and suppose that $x(\mu)$ has a limit point x_*.

Assumptions required: f and c are twice-continuously differentiable, and that $\{\nabla c_i(x_*)\}_{i \in \mathcal{A}}$ are linearly independent, where $\mathcal{A} = \{i | c_i(x_*) = 0\}$. [1 mark].

In this case

$$\nabla_x f(x(\mu)) - (\nabla c(x(\mu)))^T y(x(\mu)) = 0 \qquad \text{[1 mark]}$$

Need linear independence to ensure that $y(x(\mu)$ converges to some y_*,[1 mark], and hence that

$$\nabla_x f(x_*) - (\nabla c(x_*))^T y_*) = 0 \qquad \text{[1 mark]}$$

Since $c(x(\mu) > 0$ and $y(x(\mu)) > 0$, $c_(x_*) \geq 0$ and $y_* \geq 0$ [1 mark]. Finally, by definition

$$c_i(x(\mu)) y_i(x(\mu)) = \mu$$

and hence $c_i(x_*)(y_*)_i = 0$ for all i [1 mark]. Thus the conditions of part (a) are satisfied.

The values $y_i(x) = \mu / c_i(x)$ thus give Lagrange multiplier estimates when $x \to x(\mu)$ and $\mu \to 0$. [1 mark]

2(d) The dual feasibility condition is that

$$\begin{pmatrix} x_1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} y_1 = 0 \qquad \text{[1 mark]}$$

from which we deduce that $y_1 = 1$ and $x_1 = 0$. Since the Lagrange multiplier is positive, the constraint is active, and hence $x_2 = 0$. [1 mark].

2(e) The gradient and Hessian of the barrier function

$$\nabla_x \Phi(x,\mu) = \begin{pmatrix} x_1 \\ 1 - \mu/x_2 \end{pmatrix} \text{ and } \nabla_{xx}\Phi(x,\mu) = \begin{pmatrix} 1 & 0 \\ 0 & \mu/x_2^2 \end{pmatrix} \qquad \text{[1 mark]}$$

For points on the solution trajectory $x(\mu) = (0, 1/\mu)$, this is

$$\begin{pmatrix} 1 & 0 \\ 0 & 1/\mu \end{pmatrix}$$

Since the condition number of the Hessian is $1/\mu$, the Hessian becomes increasingly ill-conditioned as $\mu \to 0$, a naive interpretation of Newton's method might suggest that inaccurate Newton corrections may be impossible. [2 marks].

2(f) The Newton equations are

$$\begin{pmatrix} 1 & 0 \\ 0 & \mu/x_2^2 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = - \begin{pmatrix} x_1 \\ 1 - \mu/x_2 \end{pmatrix}$$

and hence

$$\begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} -x_1 \\ -(x_2/\mu)(x_2 - \mu) \end{pmatrix}.$$

Close to the solution trajectory x_2/μ is close to 1, and thus both components of d are small. Thus the ill-conditioning does not hurt. [3 marks].