1. Implement the steepest descent method with Armijo's backtracking strategy. See HW4Q1.m for input and output information. In fact, HW4Q1 applies your algorithm to the Rosenbrock function:
$$f(\mathbf{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2.$$
What do you observe from the output?

   Solution. The Rosenbrock function exhibits a zig-zag convergence. Each acceptable α found by Armijo's backtracking strategy is very small, around 0.02. The convergence is definitely linear as shown via the graphs in HW4Q1.m, but it converges nonetheless. Instead of solving for the minimum of $f(\mathbf{x})$, however, it is possible to minimize $f(T\mathbf{x})$ in hopes of faster convergence and then simply converting the new solution to the minimizer of the Rosenbrock by $x \leftarrow Tx$. The Rosenbrock function is so slow that the first **randomly** generated matrix T outperformed the original function. Letting $T = \begin{bmatrix} 3.57 & -1.34 \\ 2.76 & 3.03 \end{bmatrix}$ gives convergence in only 5813 iterations which is roughly a third of the original iterations needed.

2. Implement the Newton's method. Apply it to the Rosenbrock function. What do you observe?

   Solution. Newton's method provides a quadratically converging algorithm under convergence assumptions. In a conjugate gradient-like sequence of iterations, Newton's method converged in only 2 iterations. This is certainly much faster than the aforementioned GD approach, but did require Hessian computations.

3. Suppose that we apply Newton's method to the 1-dimensional problem
$$\min \; f(x) := tx - \ln x,$$
where $t > 0$ is a parameter. For this specific example, show that Newton's method, with starting point x_0, converges quadratically if $|x_0 - x^*| < \frac{1}{t}$ and does not converge if $|x_0 - x^*| \geq \frac{1}{t}$.

   Proof. First note that since $f''(x) = \frac{1}{x^2} \geq 0$, f is convex. Since $f'(\frac{1}{t}) = t - t = 0$, $x^* = \frac{1}{t}$ must be the global minimizer of f. Consider the absolute backward error at step $k+1$, $e_{k+1} = |x_{k+1} - x^*|$. Using the Newton method update, this is simply
$$\begin{aligned} e_{k+1} &= |x_{k+1} - x^*| \\ &= |x_k - x^* - x_k^2(t - \frac{1}{x_k})| \\ &= |2x_k - x^* - tx_k^2| \\ &= t|tx_k^2 - \frac{2x_k}{t} - \frac{1}{t^2}| \\ &= t|x_k - x^*|^2 \\ &= te_k^2 \end{aligned}$$

Using this recurrence relation, the following holds true

$$te_{k+1} = (te_k)^2 = \cdots = (te_0)^{2^{k+1}}$$

Thus, when $e_0 < \frac{1}{t}$, Newton's method applied to f will converge quadratically, and when $e_0 \geq \frac{1}{t}$, it will not converge at all. $\square$

4. As discussed in class, steepest descent, Newton, and quasi-Newton methods can be unified in the form of

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k B_k \nabla f(\mathbf{x}_k).$$

In this question, we derive the rate of convergence of such a method on a convex quadratic function

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q \mathbf{x} - \mathbf{b}^T \mathbf{x},$$

where $Q \succ 0$.

(a) Compute the step size α_k assuming that exact line search is used.

Solution. Solving for α_k via exact line search is equivalent to the minimization problem $\min f(x_k - \lambda d)$ where d is the direction chosen. For this setting, this simplifies to solving

$$\alpha_k = \operatorname{argmin} f(x_k - \lambda B_k \nabla f(x_k))$$

Since this is convex optimization, this is instead solving $\theta(\lambda) := \nabla f(x_k - \lambda B_k \nabla f(x_k)) = (B_k \nabla f(x_k))^T \nabla f(x_k - \lambda B_k \nabla f(x_k)) = 0$. Noting that $\nabla f(x_k) = Qx_k - b$, it follows that

$$\begin{aligned}
(B_k \nabla f(x_k))^T \nabla f(x_k - \lambda B_k \nabla f(x_k)) &= 0 \\
(B_k \nabla f(x_k))^T [Q(x_k - \lambda B_k \nabla f(x_k)) - b] &= 0 \\
(B_k \nabla f(x_k))^T [Q(x_k) - b] &= \lambda \nabla f(x_k)^T B_k^T Q B_k \nabla f(x_k)
\end{aligned}$$

Thus,

$$\alpha_k = \frac{(B_k \nabla f(x_k))^T \nabla f(x_k)}{\nabla f(x_k)^T B_k^T Q B_k \nabla f(x_k)}$$

is the optimal step size.

(b) Let $\mathbf{x}^*$ be the unique minimizer of f, and define $E(\mathbf{x}) := \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T Q (\mathbf{x} - \mathbf{x}^*)$. Prove that

$$E(\mathbf{x}_{k+1}) \leq \left(\frac{A_k - a_k}{A_k + a_k}\right)^2 E(\mathbf{x}_k),$$

where A_k and a_k are the largest and smallest eigenvalues of $B_k Q$, respectively.

Proof. It should be noted that the proof is nearly identical to that of the notes. The result and proof are only slightly modified. For brevity, let $A_k = B_k \nabla f(x_k)$. By expansion of

$E(x_{k+1})$ and definition of α_k,

$$\begin{aligned}
E(_{k+1}) &= \frac{1}{2}(x_{k+1} - x^*)^T Q(x_{k+1} - x^*) \\
&= \frac{1}{2}(x_k - \alpha_k A_k - x^*)^T Q(x_k - \alpha_k A_k - x^*) \\
&= \frac{1}{2}\left((x_k - x^*)^T Q(x_k - x^*) + \alpha_k^2 (A_k^T Q A_k)\right) - (\alpha_k A_k)^T Q(x_k - x^*) \\
&= -\frac{1}{2}\frac{A_k^T A_k}{A_k^T Q A_k}
\end{aligned}$$

Also not unlike the notes, it follows from definition of f that $x_k - x^* = Q^{-1}\nabla f(x_k)$ and therefore $E(x_k) = \frac{1}{2}\nabla f(x_k)^T Q^{-1} \nabla f(x_k)$. Thus,

$$E(x_{k+1}) = E(x_k)\left(1 - \frac{\nabla f(x_k)^T (B_k^{\frac{1}{2}})^T B_k^{\frac{1}{2}} \nabla f(x_k)}{\nabla f(x_k)^T B_k^T Q B_k \nabla f(x_k)(\nabla f(x_k)^T Q^{-1} \nabla f(x_k)}\right)$$

Setting $x = \nabla f(x_k)^T B_k^{\frac{1}{2}}$, we have

$$E(x_{k+1}) = E(x_k)\left(1 - \frac{x^T x}{(x^T B_k^{\frac{1}{2}} Q B_k^{\frac{1}{2}} x)(x^T B_k^{-\frac{1}{2}} Q^{-1} B_k^{-\frac{1}{2}} x)}\right)$$

Thus, Kantorovich's inequality gives us

$$E(x_{k+1}) \leq E(x_k)(1 - \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2}) = E(x_k)\left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2$$

where λ_1, λ_n are the largest and smallest eigenvalues of $B_k^{\frac{1}{2}} Q B_k^{\frac{1}{2}}$. Denote this matrix A. Since $B_k^{\frac{1}{2}} A B_k^{-\frac{1}{2}} = B_k Q$, the matrices A and $B_k Q$ are similar, and thus have the same eigenvalues. The statement follows immediately.

□

5. In the DFP method,

$$B_{k+1} = \left(I_n - \frac{\mathbf{y}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right) B_k \left(I_n - \frac{\mathbf{s}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right) + \frac{\mathbf{y}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}.$$

Prove that if $B_k \succ 0$ and $\mathbf{y}_k^T \mathbf{s}_k > 0$, then $B_{k+1} \succ 0$.

Proof. Consider the quantity $x^T B_{k+1} x$. By the DFP method, this is simply

$$x^T B_{k+1} x = x^T \left(I_n - \frac{\mathbf{y}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right) B_k \left(I_n - \frac{\mathbf{s}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right) x + x^T \left(\frac{\mathbf{y}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right) x$$

Set $A = \left(I_n - \frac{\mathbf{y}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right)$ and consider the first term. Since $A^T = \left(I_n - \frac{\mathbf{s}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right)$, the first term is $x^T A B_k A^T x = y^T B_k y$ where $y = A^T x$. Since $B_k \succ 0, y^T B_k y \geq 0$. For the second term, note that

$$x^T \left(\frac{\mathbf{y}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right) x = \left(\frac{x^T \mathbf{y}_k \mathbf{y}_k^T x}{\mathbf{y}_k^T \mathbf{s}_k}\right) = \left(\frac{\|\mathbf{y}_k^T x\|^2}{\mathbf{y}_k^T \mathbf{s}_k}\right) \geq 0$$

since $y_k^T s_k \geq 0$. Thus,

$$x^T \left(I_n - \frac{\mathbf{y}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right) B_k \left(I_n - \frac{\mathbf{s}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right) x + x^T \left(\frac{\mathbf{y}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right) x \geq 0$$

for all x and so $B_{k+1} \succ 0$. $\square$