1. (a) Let $\langle\cdot,\cdot\rangle$ be an inner product defined on a vector space V. Prove the Cauchy-Schwarz inequality

$$|\langle u,v\rangle|^2 \le \langle u,u\rangle\langle v,v\rangle \quad \forall u,v \in V. \tag{1}$$

   Hint: Consider the quadratic polynomial

$$p(t) = \langle u+tv, u+tv\rangle, \quad t \in \mathbb{R} \ (\text{or } \mathbb{C}). \tag{2}$$

   Can this polynomial take negative values? Use your answer to conclude what should be the sign of the discriminant if you set $p = 0$.

Solution: Let $u,v \in V$. Since inner products are non-negative by definition, we have

$$p(t) = \langle u+tv, u+tv\rangle \ge 0$$

for all $t \in \mathbb{R}$. Further, using the fact that inner products in linear in their first argument and (conjugate) symmetric,

$$p(t) = \langle u+tv, u+tv\rangle = \langle u, u+tv\rangle + t\langle v, u+tv\rangle$$
$$= \langle u+tv, u\rangle + t\langle u+tv, v\rangle = \langle u,u\rangle + t\langle v,u\rangle + t\langle u,v\rangle + t^2\langle v,v\rangle$$
$$= \langle u,u\rangle + 2t\langle u,v\rangle + t^2\langle v,v\rangle.$$

Therefore, we have that

$$\langle u,u\rangle + 2t\langle u,v\rangle + t^2\langle v,v\rangle \ge 0$$

for all $t \in \mathbb{R}$. Hence, in particular it must be true for

$$t^* := -\frac{\langle u,v\rangle}{\langle v,v\rangle},$$

and so

$$0 \le p(t^*) = \langle u,u\rangle - 2\frac{\langle u,v\rangle^2}{\langle v,v\rangle} + \frac{\langle u,v\rangle^2}{\langle v,v\rangle} = \langle u,u\rangle - \frac{\langle u,v\rangle^2}{\langle v,v\rangle}.$$

Finally, adding $\frac{\langle u,v\rangle^2}{\langle v,v\rangle}$ and multiplying by $\langle v,v\rangle$ on both sides yields the Cauchy-Schwarz inequality

$$\langle u,v\rangle^2 \le \langle u,u\rangle\langle v,v\rangle$$

as desired.

   (b) Let B be a real symmetric positive definite $n \times n$ matrix. Use the Cauchy-Schwarz inequality to prove that

$$(g^\top B g)(g^\top B^{-1} g) \ge (g^\top g)^2 \quad \forall g \in \mathbb{R}^n. \tag{3}$$

Solution: We first rewrite the right-hand side as a B-inner product, noting that

$$(g^\top g)^2 = (g^\top B B^{-1} g)^2 = |\langle g, B^{-1} g\rangle_B|^2$$

and so by the Cauchy-Schwarz inequality,

$$(g^\top g)^2 = |\langle g, B^{-1} g\rangle_B|^2 \le \langle g, g\rangle_B \langle B^{-1} g, B^{-1} g\rangle_B$$
$$= (g^\top B g)((B^{-1}g)^\top B(B^{-1}g)) = (g^\top B g)(g^\top B^{-1} g)$$

as desired.

2. Consider Newton's algorithm for solving the trust-region subproblem ([NW], Algorithm 4.3, page 87). Prove that Eq. (4.43) is equivalent to Eq. (4.44) in [NW], i.e., that for

$$\phi(\lambda) = \frac{1}{\Delta} - \left[\sum_{j=1}^n \frac{(q_j^\top g)^2}{(\lambda_j + \lambda)^2}\right]^{-1/2},$$

where (q_j, λ_j) are the eigenpairs of B, the Newton iteration

$$\lambda^{(l+1)} = \lambda^{(l)} - \frac{\phi(\lambda^{(l)})}{\phi'(\lambda^{(l)})}$$

is given by

$$\lambda^{(l+1)} = \lambda^{(l)} + \left(\frac{\|p_l\|}{\|z_l\|}\right)^2 \frac{\|p_l\| - \Delta}{\Delta},$$

where $z_l = L^{-1} p_l$, $p_l = -(B + \lambda^{(l)} I)^{-1} g$, and L is the Cholesky factor of $B + \lambda^{(l)} I$, i.e., $B + \lambda^{(l)} I = LL^\top$. Note: $R = L^\top$ in Algorithm 4.3.

Hint: You will need to compute the derivative of ϕ and express it in terms of $\|p_l\|$ and $\|(B + \lambda^{(l)} I)^{-1} g\|^2$. Also, you will need to use the fact that the Cholesky factor of any SPD matrix M is related to $M^{1/2}$ via an orthogonal transformation.

Solution: We first claim that
$$p_l = -\sum_{j=1}^n \frac{q_j^\top g}{\lambda_j + \lambda} q_j.$$

Since B is symmetric, it has a spectral decomposition $B = Q\Lambda Q^\top$. In particular, we can choose one such that its eigenvectors $\{q_j\}_{j=1}^n$ are orthonormal. Then

$$B + \lambda^{(l)} I = Q\Lambda Q^\top + \lambda^{(l)} I = Q\Lambda Q^\top + \lambda^{(l)} Q I Q^\top = Q(\Lambda + \lambda^{(l)} I)Q^\top$$

and hence

$$p_l = -(B + \lambda^{(l)} I)^{-1} g = -\left(Q(\Lambda + \lambda^{(l)} I)Q^\top\right)^{-1} g = -Q(\Lambda + \lambda^{(l)} I)^{-1} Q^\top g$$

meaning

$$p_l = -\sum_{j=1}^n \frac{q_j^\top g}{\lambda_j + \lambda} q_j.$$

Then since the eigenvectors $\{q_j\}_{j=1}^n$ form an orthonormal basis, this means that

$$\|p_l\| = \left[\sum_{j=1}^n \frac{(q_j^\top g)^2}{(\lambda_j + \lambda)^2}\right]^{1/2}.$$

We now claim that

$$\|z_l\| = \left[\sum_{j=1}^n \frac{(q_j^\top g)^2}{(\lambda_j + \lambda)^3}\right]^{1/2}.$$

From the hint, we know that L, the Cholesky factor of $B + \lambda(l)I$, can be written as

$$L = Q\sqrt{\Lambda + \lambda^{(l)}I}Q^\top \tilde{Q}$$

for some orthogonal matrix $\tilde{Q}$. Hence, $L^{-1} = \tilde{Q}^\top Q(\Lambda + \lambda^{(l)}I)^{-1/2}Q^\top$ and so

$$z_l = L^{-1}p_l = (\tilde{Q}^\top Q(\Lambda + \lambda^{(l)}I)^{-1/2}Q^\top)(-Q(\Lambda + \lambda^{(l)}I)^{-1}Q^\top g)$$
$$= -\tilde{Q}^\top Q(\Lambda + \lambda^{(l)}I)^{-3/2}Q^\top g.$$

Toward the desired result, recall that multiplication by an orthogonal matrix does not change a vector's norm; let Q be an orthogonal matrix and $x \in \mathbb{R}^n$. Then

$$\|Qx\|^2 = (Qx)^\top(Qx) = x^\top Q^\top Qx = x^\top x = \|x\|^2.$$

From this fact, we have that

$$\|z_l\| = \|\tilde{Q}^\top Q(\Lambda + \lambda^{(l)}I)^{-3/2}Q^\top g\| = \|Q(\Lambda + \lambda^{(l)}I)^{-3/2}Q^\top g\|$$
$$= \left\|\sum_{j=1}^n \frac{q_j^\top g}{(\lambda_j + \lambda)^{3/2}}q_j\right\| = \left[\sum_{j=1}^n \frac{(q_j^\top g)^2}{(\lambda_j + \lambda)^3}\right]^{1/2}.$$

Hence, we have shown that both

$$\|p_l\| = \left[\sum_{j=1}^n \frac{(q_j^\top g)^2}{(\lambda_j + \lambda)^2}\right]^{1/2} \text{ and } \|z_l\| = \left[\sum_{j=1}^n \frac{(q_j^\top g)^2}{(\lambda_j + \lambda)^3}\right]^{1/2}.$$

We will now compute $\phi'(\lambda)$ and write it in terms of $\|p_l\|$ and $\|z_l\|$, noting that

$$\phi'(\lambda) = \frac{d}{d\lambda}\left[\frac{1}{\Delta} - \left[\sum_{j=1}^n \frac{(q_j^\top g)^2}{(\lambda_j + \lambda)^2}\right]^{-1/2}\right] = \frac{1}{2}\left[\sum_{j=1}^n \frac{(q_j^\top g)^2}{(\lambda_j + \lambda)^2}\right]^{-3/2}\frac{d}{d\lambda}\sum_{j=1}^n \frac{(q_j^\top g)^2}{(\lambda_j + \lambda)^2}$$

$$= -\left[\sum_{j=1}^n \frac{(q_j^\top g)^2}{(\lambda_j + \lambda)^2}\right]^{-3/2}\sum_{j=1}^n \frac{(q_j^\top g)^2}{(\lambda_j + \lambda)^3} = -\|p_l\|^{-3}\|z_l\|^2.$$

Finally, we also see that

$$\phi(\lambda) = \frac{1}{\Delta} - \left[\sum_{j=1}^{n} \frac{(q_j^\top g)^2}{(\lambda_j + \lambda)^2} \right]^{-1/2} = \frac{1}{\Delta} - \frac{1}{\|p_l\|},$$

and hence we have

$$-\frac{\phi(\lambda^{(l)})}{\phi'(\lambda^{(l)})} = \frac{\frac{1}{\Delta} - \frac{1}{\|p_l\|}}{\|p_l\|^{-3}\|z_l\|^2} = \frac{\frac{\|p_l\|^3}{\Delta} - \|p_l\|^2}{\|z_l\|^2} = \frac{\|p_l\|^2}{\|z_l\|^2} \frac{\frac{\|p_l\|}{\Delta} - 1}{1} = \left(\frac{\|p_l\|}{\|z_l\|} \right)^2 \frac{\|p_l\| - \Delta}{\Delta}$$

meaning that Eq. (4.43) and Eq. (4.44) are indeed equivalent.

3. Consider the problem of finding local energy minima of the LJ_7 as in Problem 3 of HW9. Consider the same set of initial conditions: four initial conditions close to its four local minima, and ten random initial conditions.

Implement the BFGS trust-region method with the dogleg subproblem solver. Compare its performance with the trust-region Newton with the exact subproblem solver implemented in the provided code by creating a table with the number of iterations required to achieve convergence and plotting the graph of f and $\|\nabla f\|$ against the iteration number for each test case (the four initial conditions close to the minima and one representative random configuration initial condition). Do it for each of the four initial conditions approximating the four local minima and ten random initial conditions. The set of figures to include is the same as for Problem 3 in HW9.

Comment on the performance of trust-region methods compared to the performance of line-search methods.

Solution: The code and plots for all fourteen runs can be found here. We supply the plots for model 1 below.

Lennard-Jones Trust Region with Model 1
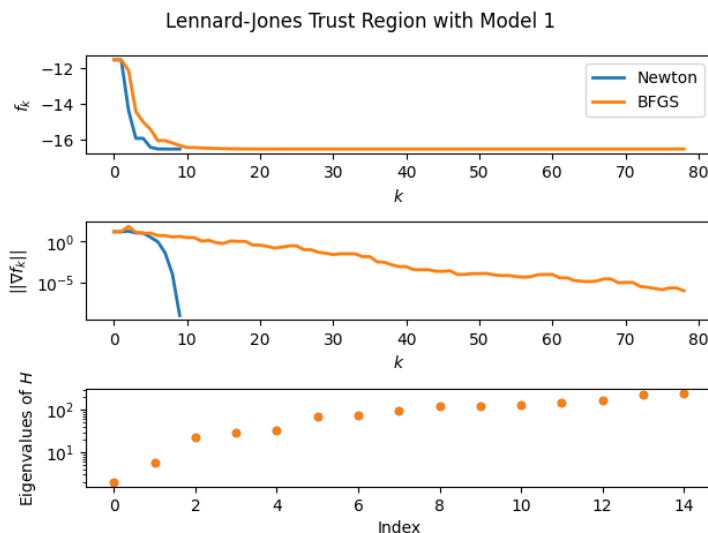
> We see visually that the Newton's method convergences faster than the BFGS method. The converges is comparable (possibly slightly faster) than that of line search.

4. (Approx. Problem 3.1 from [NW]) Write a code that applies the two algorithms from the previous problem (the trust-region BFGS with the dogleg solver and the trust-region Newton with the exact subspace solver) to the Rosenbrock function as in Problem 4 of HW9:

$$f(x, y) = 100(y - x^2)^2 + (1 - x)^2. \tag{4}$$

Experiment with the same two initial conditions: $(1.2, 1.2)$ and $(-1.2, 1)$.

Plot the level sets of the Rosenbrock function using the command contour and plot the iterations for each method over it. Plot $\|(x_k, y_k) - (x^*, y^*)\|$ versus k in the logarithmic scale along the y-axis for each method. Compare the performance of the methods.

Solution: We plot the figure for the harder intitial condition $(-1.2, 1)$ below. The plots for the other initial condition can be found at here.
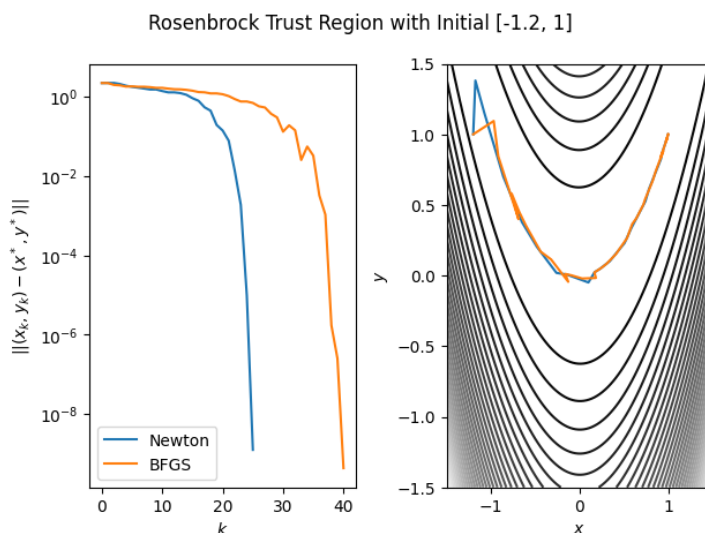
As with the previous question, Newton's method converges in approximately 25 iterations, faster than that BFGS, which converges in approximately 40 iterations.