Exercise 1. A financial problem.

Problem Statement.
We wish to consider maximizing an earnings function \( E(y, z) \), where the variable \( y \) represents selling price and the variable \( z \) represents advertising expenses. Although in practice we can only deal with \( y \geq 0 \) and \( z \geq 0 \) we will treat this as an unconstrained problem. The form of the earnings function is 
\[
E = yx - (z + g_2(x)),
\]
where
\[
\begin{align*}
x &= g_1(y, z) = a_1 + a_2 y + a_3 z + a_4 y z + a_5 z^2 \\
g_2(x) &= e_1 + e_2 x.
\end{align*}
\]

The values of the constants are \( a_1 = 50000, a_2 = -5000, a_3 = 40, a_4 = -1, a_5 = -0.02, e_1 = 100000, \) and \( e_2 = 2 \). A plot of the function is shown in Figure 1.

Preliminary Analysis.
Define \( \hat{E} = -E \). After some simple manipulations, we find that
\[
\hat{E} = A + By + Cz + Dy^2 + Fz^2 + Gyz + Hy^2 z + Jyz^2,
\]
where
\[
\begin{align*}
A &= e_1 + e_2 a_1 \\
B &= e_2 a_2 - a_1 \\
C &= 1 + a_3 e_2 \\
D &= -a_2 \\
F &= e_2 a_5 \\
G &= e_2 a_4 - a_3 \\
H &= -a_4 \\
J &= -a_5.
\end{align*}
\]
We will also later need the gradient and Hessian of $\hat{E}$. We find
\[
\nabla \hat{E} = \begin{bmatrix}
B + 2Dy + Gz + 2Hzy + J z^2 \\
C + 2Fz + Gy + H y^2 + 2J yz
\end{bmatrix}
\]
\[
\nabla^2 \hat{E} = \begin{bmatrix}
2D + 2Hz & G + 2Hz + 2J z \\
G + 2Hy + 2J z & 2F + 2J y
\end{bmatrix}
\]

Implementation Overview.
The following methods have been implemented:

- Pure steepest descent
- Pure Newton’s method
- Newton’s method with Cholesky factorization
- Scaled steepest descent.

In addition, the following step sizes have been tested:

- Constant
- Armijo rule
• Line minimization using golden section search and parabolic interpolation (for Newton w/ Cholesky).

In all the methods, the convergence criterion was \( \|\nabla f(x^k)\|/\|\nabla f(x^0)\| < 10^{-10} \). The optimal solution is (roughly) \((y^*, z^*) \approx (10.559, 7330.9)\).

**Computational Results.**

1. **Pure steepest descent.**
   Without scaling, steepest descent is not well-suited for this problem at all. We observe that, even from a good starting point, we have not converged to the optimal solution after one million iterations. See Figure 2 for an illustration.

![Figure 2: Steepest descent struggles.](image)

2. **Pure Newton’s method.**
   If we use a constant stepsize, we find that it is possible to converge to the optimal solution if we start where the Hessian is already positive definite, as in Figure 3. On the other hand, we see that the choice of the stepsize is critical, as choosing it too large can lead us off towards saddle points (see Figure 4). If we change to the Armijo rule (where in all cases, \(\sigma = 10^{-5}\) has been used), we generally see an improvement, both in ability to find the optimal solution from various starting points, and in convergence rate.
There are cases, however, for which Newton’s method with the Armijo rule fails.

![Figure 3: Close to the optimum, Newton’s method with constant stepsize worked.](image)

3. **Newton’s method with Cholesky factorization.**
   
   Here we use Newton’s method and apply modified Cholesky factorization to the Hessian whenever it is not positive definite. We use the Armijo rule when the Hessian is unchanged and a line minimization search otherwise. When starting in a region near the optimal solution (i.e., the Hessian is positive definite), this method behaves similarly to the unmodified Newton’s method with Armijo. For more difficult points, however, this method produced convergence, as in Figure 5. Still, some starting points gave even this method difficulties (Figure 6).

4. **Scaled steepest descent.**
   
   In this method we used a scaling matrix $D_k$ at each iteration which was the inverse of the diagonal terms of the Hessian evaluated at $x^k$ (with the modification that if any of these diagonal terms were less than $10^{-6}$, they were replaced by unity). First, we see that where the pure steepest descent took more than one million iterations, we converge here in 16 iterations. We are also able to converge from challenging initial guesses (such as where modified Newton did not). As with all other implementations, however, we
found starting points which made the method choke. Figure 7 highlights one succesful run with this method.

**Summary of Results.**
Clearly, pure steepest descent is not a good choice for this problem. For Newton’s method, we have found that it is quite effective when used with modified Cholesky factorization. Likewise, diagonally scaled steepest descent worked well. On the other hand, no method guaranteed convergence to the optimal solution from all starting points.

**Exercise 2.** *Exercise 1.5.3.*

**Preliminary Analysis.**
Here we have $g : \mathbb{R}^{2} \mapsto \mathbb{R}^{5}$, where $g_i(u_0, u_1) = z_i - \tanh(u_0 + y_i u_1)$. We know that the gradient matrix of $g$ will be of size $2 \times 5$, where column $i$ of $\nabla g(u_0, u_1)$ is given by

$$\nabla g_i(u_0, u_1) = \begin{bmatrix} -\text{sech}^2(u_0 + y_i u_1) \\ -y_i \text{sech}^2(u_0 + y_i u_1) \end{bmatrix}.$$
Implementation Overview.
The following methods have been implemented:
- Pure Gauss-Newton method
- Gauss-Newton method with Cholesky factorization
- Incremental gradient method
- Incremental version of the heavy ball method.
In addition, the following step sizes have been tested:
- Constant
- Armijo rule
- Line minimization (for Newton w/ Cholesky).
- Diminishing stepsizes.
In all the methods, the convergence criterion was $\|\nabla f(x^k)\|/\|\nabla f(x^0)\| < 10^{-10}$. The optimal solution is (roughly) $(u_0^*, u_1^*) \approx (.2261, -.0806)$.

Computational Results.
1. **Pure Gauss-Newton method.**
   This method generally works well starting close to the optimal solution. Unfortunately, very often this method does not converge. Figure 9 and 10 demonstrate the inconsistencies here.

2. **Gauss-Newton method with Cholesky factorization.**
   This method uses modified Cholesky factorization to ensure that the Hessian is always positive definite. We use the Armijo rule on iterations when the Hessian is unchanged, and a line minimization routine otherwise. The behavior of this modified G-N method is clearly better than its original counterpart. We converge here for cases where we did not converge before, and the number of iterations to convergence is typically smaller. Still, there are cases which do not result in convergence.

3. **Incremental gradient method.**
   If we use a constant stepsize of value 0.2, we converge towards a limit point, which, although close to the optimum, differs from the minimum. With a diminishing stepsize, however, the method converges to the minimum from these same starting points. Note that this method tends to get us “close” to the optimal point quickly, but near the optimum, progress gets rapidly slower.
Figure 7: Steepest descent worked much more effectively with scaling.

4. **Incremental very of the heavy ball method.**
   The heavy ball method did not seem to work particularly well overall. We see that with both a constant and a diminishing stepsize, we can converge to limit points which are not minima. As the momentum term $\beta$ is reduced closer towards 0, the limit point becomes closer and closer to the true optimum. Figure 15 and 16 highlight this.

**Summary of Results.**
The least squares function we are dealing with is quite difficult to optimize. Clearly, modifying Gauss-Newton with Cholesky factorization results in an improvement over the pure Gauss-Newton method, and we do have cases for both which result in convergence. The incremental gradient method is also satisfactory for some starting points, particularly when combined with a diminishing stepsize. Using a constant stepsize here results in limit points which are not at the minimum. The incremental version of the heavy ball method did not perform well in the tests here.
Figure 8: The isocurves of the function.
Figure 9: Pure Gauss-Newton worked well near the optimum.
Figure 10: Some relatively close points cause pure GN to fail as well.
Figure 11: Modified Gauss-Newton could handle cases which caused the pure method to fail.
Figure 12: Initial points somewhat far from the optimum caused modified GN to fail.
Figure 13: Inc. gradient method with constant stepsize converged to a point near but not the optimum.
Figure 14: Inc. gradient method using diminishing stepsize converged correctly.
Figure 15: The heavy ball method did not seem to work well here.
Figure 16: Heavy ball performance improved with a smaller $\beta$. 