DSA3102: Solutions to Tutorial Set 13
Assigned: 09/11/23

1. BV Problem 9.9

   Solution: The first expression follows from a change of variables

   $$w = \nabla^2 f(x)^{1/2} v, \qquad v = \nabla^2 f(x)^{-1/2} w$$

   and from

   $$\sup_{\|w\|_2=1} -w^T \nabla^2 f(x)^{-1/2} \nabla f(x) = \|\nabla^2 f(x)^{-1/2} \nabla f(x)\|_2 = \lambda(x)$$

   The second expression follows immediately from the first.

2. BV Problem 9.10

   Solution:

   - $f(x) = \log(e^x + e^{-x})$ is a smooth convex function, with a unique minimum at the origin. The pure Newton method started at $x^{(0)} = 1$ produces the following sequence.

k	$x^{(k)}$	$f(x^{(k)}) - p^\star$
1	$-8.134 \cdot 10^{-01}$	$4.338 \cdot 10^{-1}$
2	$4.094 \cdot 10^{-01}$	$2.997 \cdot 10^{-1}$
3	$-4.730 \cdot 10^{-02}$	$8.156 \cdot 10^{-2}$
4	$7.060 \cdot 10^{-05}$	$1.118 \cdot 10^{-3}$
5	$-2.346 \cdot 10^{-13}$	$2.492 \cdot 10^{-9}$

     Started at $x^{(0)} = 1.1$, the method diverges.

k	$x^{(k)}$	$f(x^{(k)}) - p^\star$
1	$-1.129 \cdot 10^{0}$	$5.120 \cdot 10^{-1}$
2	$1.234 \cdot 10^{0}$	$5.349 \cdot 10^{-1}$
3	$-1.695 \cdot 10^{0}$	$6.223 \cdot 10^{-1}$
4	$5.715 \cdot 10^{0}$	$1.035 \cdot 10^{0}$
5	$-2.302 \cdot 10^{4}$	$2.302 \cdot 10^{4}$

   - $f(x) = -\log x + x$ is smooth and convex on **dom**, $f = \mathbb{R}_{++}$, with a unique minimizer at $x = 1$. The pure Newton method started at $x^{(0)} = 3$ gives as first iterate

     $$x^{(1)} = 3 - \frac{f'(3)}{f''(3)} = -3$$

     which lies outside **dom** f.

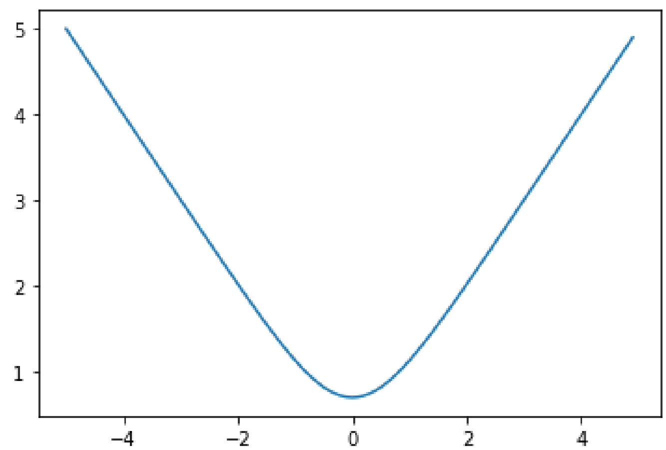
   The code can be found here.

In [7]:
```
import numpy as np
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

In [8]:
```
def f(x):
    return np.log(np.exp(x)+np.exp(-x))

def f_prime(x):
    return 1-2*np.exp(-x)/(np.exp(x)+np.exp(-x))

def f_double_prime(x):
    return 4*np.exp(2*x) / ((np.exp(2*x)+1)**2)
```

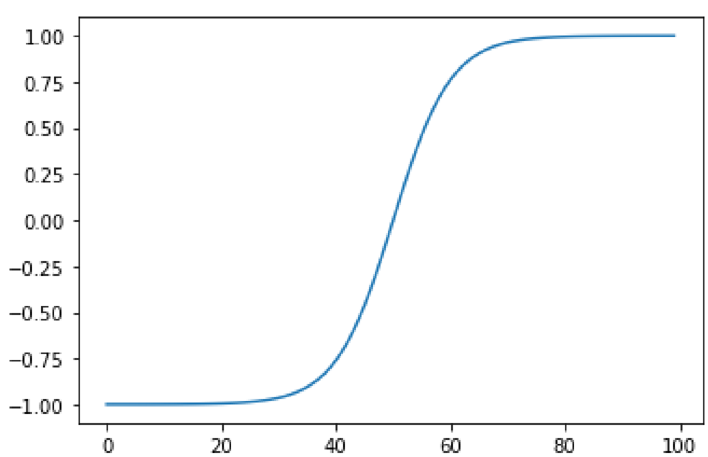
Now we first plot f, together with its first order derivative f'

In [9]:
```
x = [-5+i*0.1 for i in range(100)]
y = [f(item) for item in x]
y_prime = [f_prime(item) for item in x]

plt.plot(x,y)
plt.show()
```



Clearly, we obtain the global minimum at $x^{\star}=0$

In [10]:
```
plt.plot(y_prime)
plt.show()
```



Now we implement pure Newton's method with step size $t=1$ constantly. In other words, we have $x^{(k+1)} = x^{(k)} + \Delta x_{nt}$, for $k = 0,1,2,...$

In [11]:
```
x0 = 1
x0_alternative = 1.1

x_k = [x0]
y_k = [x0_alternative] # We define two trajectories with distinct starting po
for i in range(5): # We perform 5 iterations of Newton's method. Thus, in thi
    delta_x_nt = -1/f_double_prime(x_k[i])*f_prime(x_k[i])
    delta_y_nt = -1/f_double_prime(y_k[i])*f_prime(y_k[i])

    x_k.append(x_k[i]+delta_x_nt)
    y_k.append(y_k[i]+delta_y_nt)

p_star = [f(0) for i in range(6)]

optimality_gap_x0 = list(np.array([f(item) for item in x_k]) - np.array(p_sta
optimality_gap_x0_alternative = list(np.array([f(item) for item in y_k]) - np

print("If we start from x0=1, this is what happens to the optimality gap for 
print('\n')
print("If we start from x0=1.1, this is what happens to the optimality gap fo
print('\n')
```

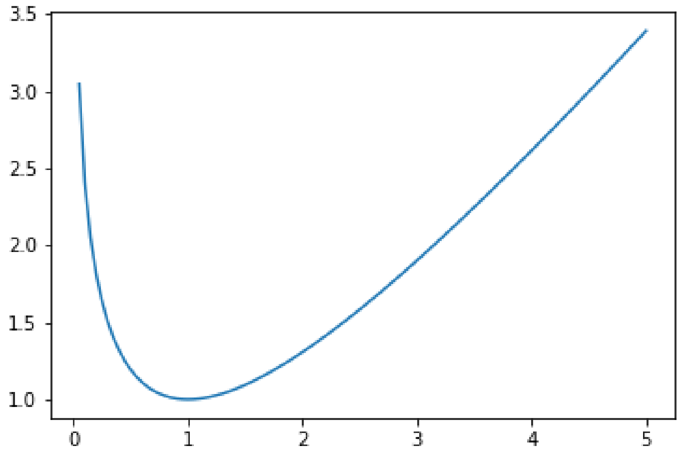
```
If we start from x0=1, this is what happens to the optimality gap for each ro
und: [0.4337808304830272, 0.2997218287983928, 0.08156361618530028, 0.00111846
05136171921, 2.492377859653061e-09, 0.0]


If we start from x0=1.1, this is what happens to the optimality gap for each
round: [0.5119361392087508, 0.534936662546477, 0.6223168792455797, 1.03516096
8649203, 5.022223776547119, inf]
```

Now we repeat the same procedure for part b)

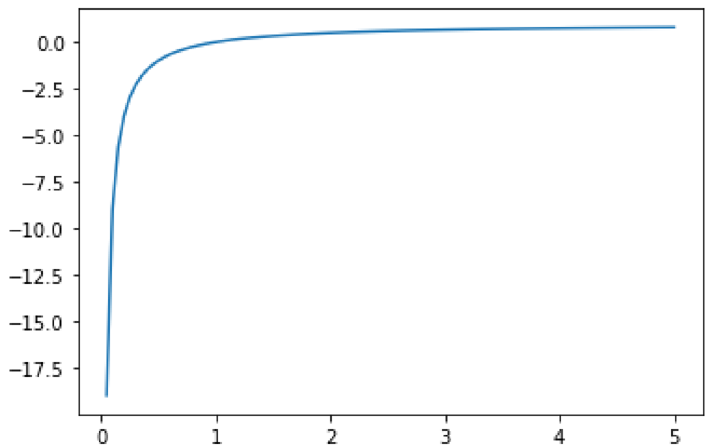
In [12]:
```
def f(x):
    return -np.log(x)+x

def f_prime(x):
    return -1/x + 1

def f_double_prime(x):
    return 1/(x**2)
```

Now we first plot \$f\$, together with its first order derivative \$f'\$

In [13]:
```
x = [(i+1)*0.05 for i in range(100)]
y = [f(item) for item in x]
y_prime = [f_prime(item) for item in x]

plt.plot(x,y)
plt.show()
```



In [14]:
```
plt.plot(x,y_prime)
plt.show()
```



In [15]:
```
x0 = .1

num_iter = 10

x_k = [x0]
for i in range(num_iter): # We perform 5 iterations of Newton's method. Thus,
    delta_x_nt = -1/f_double_prime(x_k[i])*f_prime(x_k[i])

    x_k.append(x_k[i]+delta_x_nt)

p_star = [f(1) for i in range(num_iter+1)]

optimality_gap_x0 = list(np.array([f(item) for item in x_k]) - np.array(p_sta

print("If we start from x0=3, this is what happens to the optimality gap for
print('\n')
```

```
If we start from x0=3, this is what happens to the optimality gap for each ro
und: [1.4025850929940455, 0.8507312068216506, 0.41130436154391625, 0.13247171
09306084, 0.019635790855335955, 0.0006033611663580629, 6.955890572424295e-07,
9.661160760288112e-13, 0.0, 0.0, 0.0]
```

Meaning that at the second step(i.e. $k=1$), $x^{(k)}$ is already outside of $Dom(f)$(which leads to the nan expression in the outputs above).

In [16]:
```
np.ones(num_iter+1)-x_k
```

Out[16]:
```
array([9.00000000e-01, 8.10000000e-01, 6.56100000e-01, 4.30467210e-01,
       1.85302019e-01, 3.43368382e-02, 1.17901846e-03, 1.39008452e-06,
       1.93234317e-12, 0.00000000e+00, 0.00000000e+00])
```

In [ ]:

3. BV Problem 9.11
4. **Solutions:**
   (a) Gradient method: The gradients are positive multiples
   $$\nabla g(x) = \phi'(f(x))\nabla f(x)$$
   so with exact line search the iterates are identical for f and g. With backtracking there can be big differences.

   (b) Newton method: The Hessian of g is
   $$\phi''(x)\nabla f(x)\nabla f(x)^T + \phi'(f(x))\nabla^2 f(x)$$
   so the Newton direction for g is
   $$-\left(\phi''(x)\nabla f(x)\nabla f(x)^T + \phi'(f(x))\nabla^2 f(x)\right)^{-1}\nabla f(x)$$
   From the matrix inversion lemma, we see that this is some positive multiple of the Newton direction for f. Hence with exact line search, the iterates are identical.

   Without exact line search, e.g., with Newton step one, there can be big differences. Take for example $f(x) = x^2$ and $\phi(x) = x^2$ for $x \geq 0$.

5. The purpose of this exercise is to show that Newton's method is unaffected by a linear scaling of the variables. Consider a linear invertible transformation of variables $x = Sy$. Write Newton's method in the space of the variables y and show that it generates the sequence $y^{(k)} = S^{-1}x^{(k)}$ where $\{x^{(k)}\}_{k\geq 1}$ is the sequence generated by Newton's method in the space of the variables x.

   Solution: Consider the problem
   $$\min_y \ h(y) = f(Sy)$$
   Newton's method generates the iterates according to
   $$y^{(k+1)} = y^{(k)} - t_k(\nabla^2 h(y^{(k)}))^{-1}\nabla h(y^{(k)})$$
   We have
   $$\nabla h(y) = S^T\nabla f(Sy), \qquad \nabla^2 h(y) = S^T\nabla^2 f(Sy)S$$
   So Newton's method in the space of y yields
   $$\begin{aligned}
   Sy^{(k+1)} &= Sy^{(k)} - t_kS(\nabla^2 h(y^{(k)}))^{-1}\nabla h(y^{(k)}) \\
   &= Sy^{(k)} - t_kS(S^T\nabla^2 f(Sy^{(k)})S)^{-1}S^T\nabla f(Sy^{(k)}) \\
   &= Sy^{(k)} - t_kSS^{-1}(\nabla^2 f(Sy^{(k)}))^{-1}(S^T)^{-1}S^T\nabla f(Sy^{(k)}) \\
   &= Sy^{(k)} - t_k(\nabla^2 f(Sy^{(k)}))^{-1}\nabla f(Sy^{(k)})
   \end{aligned}$$
   By replacing $Sy^{(k)}$ with $x^{(k)}$ we have
   $$x^{(k+1)} = x^{(k)} - t_k(\nabla^2 f(x^{(k)}))^{-1}\nabla f(x^{(k)})$$
   which is Newton's method in the space of the variables x.

6. Consider the pure (stepsize equals 1) form of the Newton method for the case of the cost function
   $$f(x) = \|x\|^\beta$$

for $\beta > 1$. For what starting points and values of β does the method converge to the optimal solution? What happens when $\beta \leq 1$?

You will find the matrix equality

$$(A + CBC^T)^{-1} = A^{-1} - A^{-1}C(B^{-1} + C^T A^{-1} C)^{-1} C^T A^{-1}$$

useful in inverting the Hessian.

Solution: We first calculate the derivatives as

$$\begin{aligned} \nabla f(x) &= \beta \|x\|^{\beta-2} x, \\ \nabla^2 f(x) &= \beta(\beta-2)\|x\|^{\beta-4} x x^T + \beta \|x\|^{\beta-2} I. \end{aligned}$$

We guess that the Newton direction has the form $d = -\gamma x$, where γ is a scalar, and we check the equation $\nabla^2 f(x) = \gamma \nabla f(x)$ to determine the appropriate value of γ. In this way, we obtain

$$\gamma = -\frac{1}{\beta - 1}.$$

We could also have used the matrix inversion formula to determine the inverse of $\nabla^2 f(x)$. Indeed, get from the derivation of the Hessian above that

$$\left(\nabla^2 f(x)\right)^{-1} = \frac{1}{\beta \|x\|^{\beta-2}} \left((\beta - 2) \frac{x x^T}{\|x\|^2} + I \right)^{-1}.$$

Using the substitutions $A = I \in \mathbb{R}^{n \times n}$, $B = \beta - 2 \in \mathbb{R}$, and $C = \frac{x}{\|x\|} \in \mathbb{R}^n$ (and so $C^T = (\frac{x}{\|x\|})^T$) in the matrix inversion formula. We obtain

$$\begin{aligned} \left(I + (\beta-2) \frac{x x^T}{\|x\|^2} \right)^{-1} &= I - \frac{x}{\|x\|} \left(\frac{1}{\beta - 2} + \left(\frac{x}{\|x\|} \right)^T \left(\frac{x}{\|x\|} \right) \right)^{-1} \left(\frac{x}{\|x\|} \right)^T \\ &= I - \left(\frac{x}{\|x\|} \right) \left(\frac{1}{\beta-2} + 1 \right)^{-1} \left(\frac{x}{\|x\|} \right)^T \\ &= I - \frac{\beta-2}{\beta-1} \left(\frac{x}{\|x\|} \right) \left(\frac{x}{\|x\|} \right)^T. \end{aligned}$$

Thus,

$$\left(\nabla^2 f(x)\right)^{-1} = \frac{1}{\beta \|x\|^{\beta-2}} \left(I - \frac{\beta-2}{\beta-1} \frac{x x^T}{\|x\|^2} \right).$$

The Newton direction is

$$\begin{aligned} -\left(\nabla^2 f(x)\right)^{-1} \nabla f(x) &= -\frac{1}{\beta \|x\|^{\beta-2}} \left(I - \frac{\beta-2}{\beta-1} \frac{x x^T}{\|x\|^2} \right) \beta \|x\|^{\beta-2} x \\ &= -\left(x - \frac{\beta-2}{\beta-1} \frac{x x^T}{\|x\|^2} x \right) \\ &= -\frac{1}{\beta-1} x. \end{aligned}$$

Thus Newton's method is

$$x^{(k+1)} = x^{(k)} - \left(\nabla^2 f(x^{(k)})\right)^{-1} \nabla f(x^{(k)}) = x^{(k)} - \frac{1}{\beta-1} x^{(k)} = \frac{\beta-2}{\beta-1} x^{(k)}.$$

Therefore

$$\|x^{(k+1)}\| = \left| \frac{\beta-2}{\beta-1} \right| \|x^{(k)}\|.$$

Hence if $\beta > 3/2$, we have $\left|\frac{\beta-2}{\beta-1}\right| < 1$ for and so the method converges to 0 for any initial point x^0. In particular, for $\beta = 2$ it converges in one step as expected. If $\beta = 3/2$, the method generates the points $x^{(k)}$ on the sphere $S = \{x : \|x\| = \|x^{(0)}\|\}$ and does not converge for any initial point $x^{(0)} \neq 0$. For $1 < \beta < 3/2$, we have $\left|\frac{\beta-2}{\beta-1}\right| > 1$ and the method diverges for all $x^{(0)} \neq 0$.

Now let's look at the case $\beta \leq 1$. Since

$$(\nabla^2 f(x))^{-1} = \frac{1}{\beta\|x\|^{\beta-2}} \left(I - \frac{\beta-2}{\beta-1} \frac{xx^T}{\|x\|^2} \right)$$

we see that $(\nabla^2 f(x))^{-1}$ does not exist for $\beta = 1$. If $\beta < 1$, then $\left|\frac{\beta-2}{\beta-1}\right| = \frac{2-\beta}{1-\beta} > 1$ and the method diverges for any initial point.

7. By making the same assumptions as in the "Analysis of Newton's method" document on the course homepage show the following:

Proposition 1. *Given any $\epsilon > 0$, there exists a $\delta > 0$ such that if $\|x^{(k)} - x^*\| < \delta$, then*

$$\|x^{(k+1)} - x^*\| \leq \epsilon\|x^{(k)} - x^*\|, \quad \textit{and} \quad \|g(x^{(k+1)})\| \leq \epsilon\|g(x^{(k)})\|$$

This exercise shows that g also decreases superlinearly when $x^{(k)}$ is sufficiently close to x^*. Recall that g corresponds to the gradient ∇f in Newton's method.

Solution: From the proof in my document, we see that

$$\|x^{k+1} - x^*\| \leq M \left(\int_0^1 \|\nabla g(x^*) - \nabla g(x^* + t(x^k - x^*))\| \, \mathrm{d}t \right) \|x^k - x^*\|$$

By continuity of ∇g, we can take δ sufficiently small to ensure that the term under the integral sign is arbitrarily small. Let δ_1 be such that the term under the integral sign is less than r/M. Then

$$\|x^{k+1} - x^*\| \leq r\|x^k - x^*\|$$

Now let

$$M(x) = \int_0^1 \nabla g(x^* + t(x - x^*)) \, \mathrm{d}t$$

We then have $g(x) = M(x)(x - x^*)$. Note that $M(x^*) = \nabla g(x^*)$. We have that $M(x^*)$ is invertible. By continuity of ∇g, we can take δ to be such that the region S_δ around x^* is sufficiently small so the matrix $M(x)^T M(x)$ is invertible. Let δ_2 be such that $M(x)^T M(x)$ is invertible. Then the eigenvalues of $M(x)^T M(x)$ are all positive. Let γ and Γ be such that

$$0 < \gamma \leq \min_{\|x-x^*\| \leq \delta_2} \lambda(M(x)^T M(x)) \leq \max_{\|x-x^*\| \leq \delta_2} \lambda(M(x)^T M(x)) \leq \Gamma$$

where $\lambda(\cdot)$ denotes the vector of eigenvalues of the said matrix. Then since

$$\|g(x)\|^2 = (x - x^*)^T M(x)^T M(x)(x - x^*),$$

we have

$$\gamma\|x - x^*\|^2 \leq \|g(x)\|^2 \leq \Gamma\|x - x^*\|^2$$

or

$$\frac{1}{\sqrt{\Gamma}}\|g(x^{k+1})\| \leq \|x^{k+1} - x^*\| \quad \text{and} \quad r\|x^k - x^*\| \leq \frac{r}{\sqrt{\gamma}}\|g(x^k)\|$$

Since we've already shown that $\|x^{k+1} - x^*\| \leq r\|x^k - x^*\|$, we have

$$\|g(x^{k+1})\| \leq \frac{r\sqrt{\Gamma}}{\sqrt{\gamma}}\|g(x^k)\|$$

Let $\hat{r} = \frac{r\sqrt{\Gamma}}{\sqrt{\gamma}}$. By letting $\hat{\delta}$ be sufficiently small, we have $\hat{r} < r$. By letting $\delta = \min\{\hat{\delta}, \delta_2\}$, we have for any r both desired results.