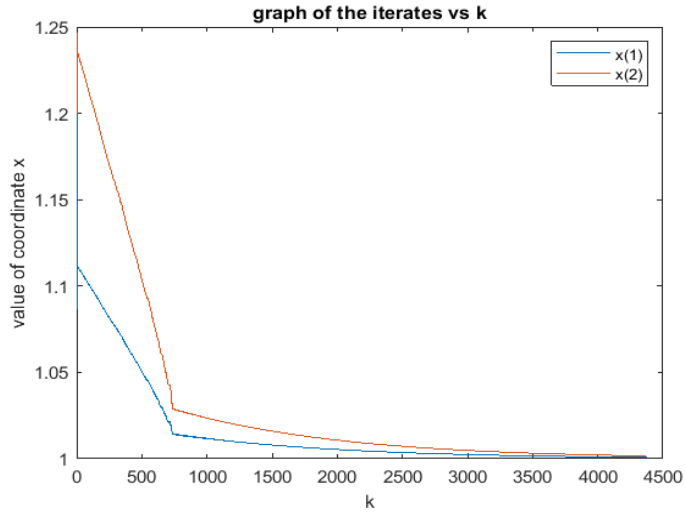
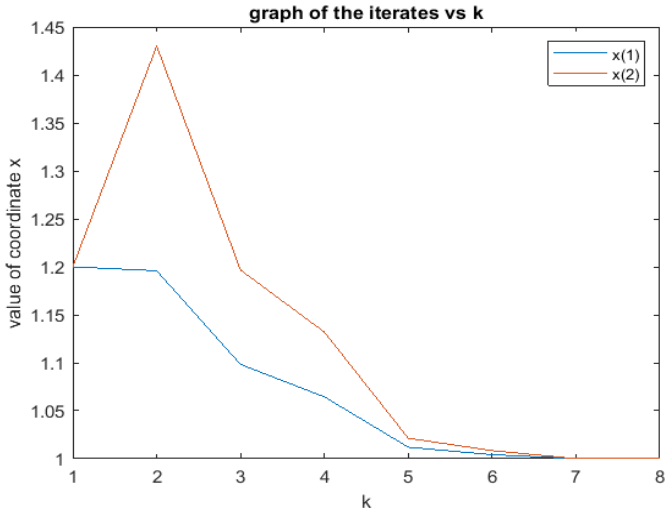
# 1 Question 1

Do Questions 3.1 and 3.9 on pages 63 and 64, respectively, of the Nocedal and Wright textbook.

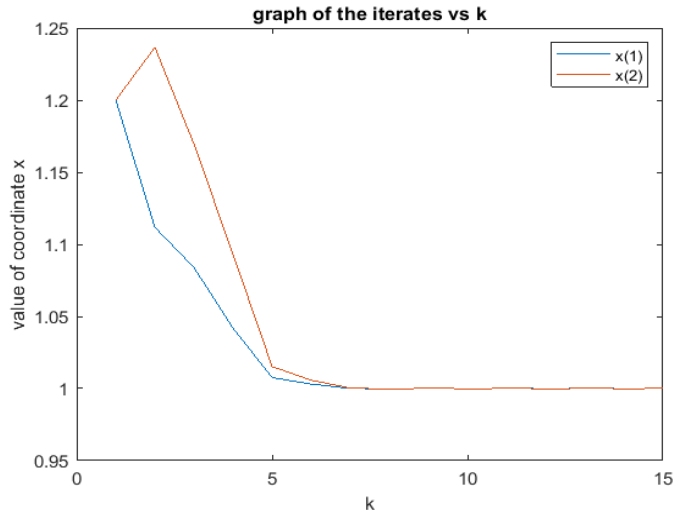
**Initial point [1.2, 1.2]**




Steepest Descent Method




Newtons Method




BFGS Method

For all the methods above the stopping criteria was chosen to be the same ($\left\|\nabla f(x_k)\right\|_2 \leq 10^{-3}$)
It is clear that all 3 methods converge to the minimum started from [1.2, 1.2]. We can see that Newtons method takes the least number of steps to converge, followed closely by BFGS and Steepest Descent taking a very large number of steps. The reason this happens is because Newtons directly exploits the second order information from computing the Hessian at each step, and BFGS does the same but instead by keeping approximations to the Hessian (less computationally expensive approach).
Now looking at the convergence ratios for the problems.
**For steepest descent** because it takes so many iterations to converge, only every 100 th line of the table is printed/recorded. The ratio $\frac{\|x_k - x^*\|_2}{\|x_{k-1} - x^*\|_2}$ seems to consistently be around 0.999 for all the iterates k. This suggests a sublinearly convergence rate to the minimum for the method. (seems like the ration tends to 1 as k keeps increasing). The ratio $\frac{\|x_k - x^*\|_2}{\|x_{k-1} - x^*\|_2^2}$ does not appear to be bounded and just keeps increasing. This is no surprise since we do not expect steepest descent to have a quadratic convergence.

**For newtons method** there are not a lot of table entries since it converges so fast. But what we do see is suggests

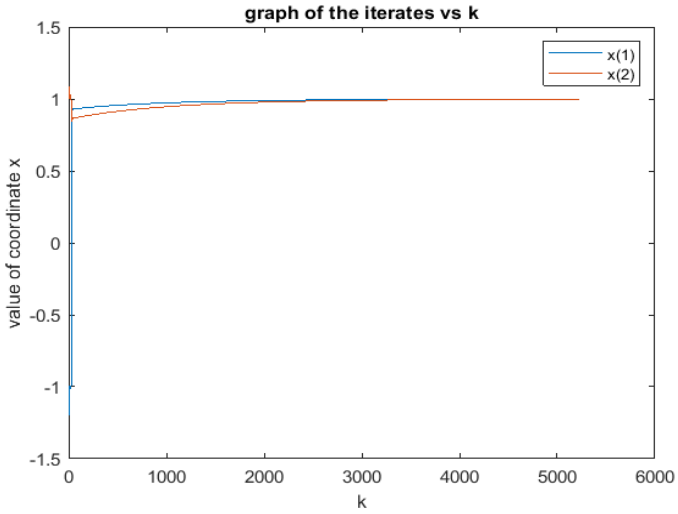
that this method has superlinear converges (ratio $\frac{\|x_k - x^*\|_2}{\|x_{k-1} - x^*\|_2}$ appears to go to 0 as k increases) as well it looks like the method actually has quadratic convergence. The ratio $\frac{\|x_k - x^*\|_2}{\|x_{k-1} - x^*\|_2^2}$ does not appear to become very large (except at k = 5 it becomes a bit larger). This agrees with Theory that newtons method has quadratic convergence started close enough to the minimum (which is the case here).

**For BFGS** the ratios $\frac{\|x_k - x^*\|_2}{\|x_{k-1} - x^*\|_2}$ suggest that it definitely has linear convergence because they appear to be bounded by 0.6 for the larger k. Although theory says the convergence rate for Quasi-Newton methods is often superlinear, the results here are inconclusive to support that claim.

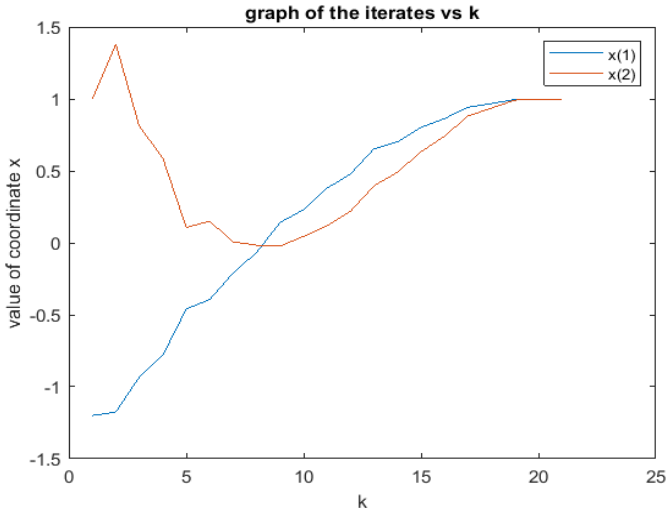
For initial point [-1.2, 1]
For this starting point all methods take a little longer to converge to the minimum than they did starting from point [1, 1]. Steepest descent takes about 1000 more steps and interestingly BFGS performs almost as well as Newtons, both take about 20 steps.

The results for convergence rates are virtually the same as for starting point [1, 1]. Steepest descent convergence appears to still be sublinear. BFGS results are still inconclusive on whether or not it is superlinear but definitely at least linear. Newtons still appears to converge quadratically to the minimum even from this point (suggests it is still started 'close enough' to a minimum)

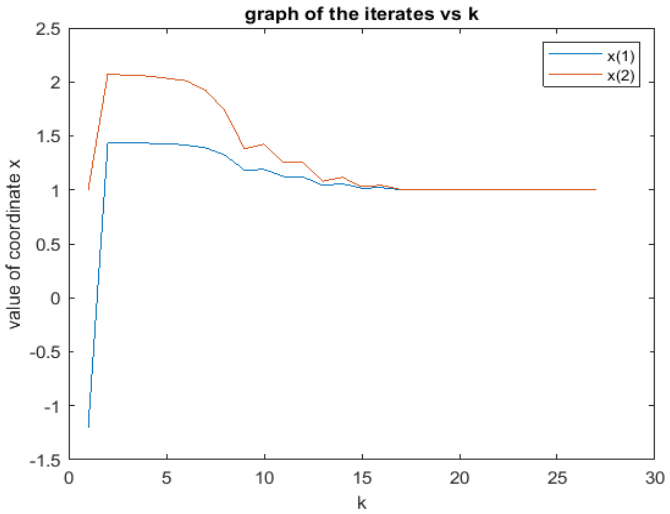
**Initial point [-1.2, 1]**




Steepest Descent Method

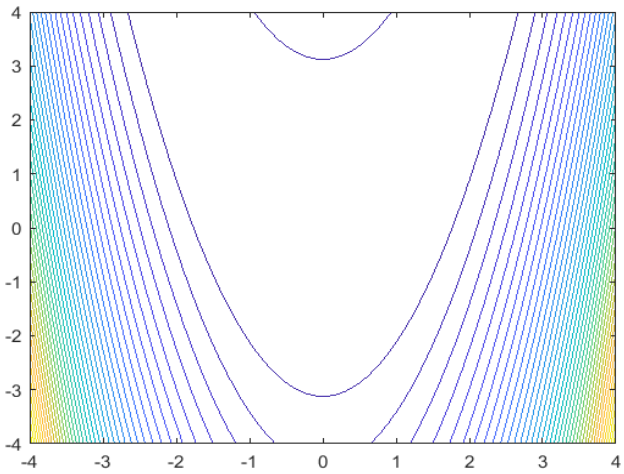



Newtons Method




BFGS Method

**Contour**



A few level sets of Rosenbrock's function

## 2 Question 2

Do Question 3.7 on page 64 of the Nocedal and Wright textbook. Show

$$\|x_{k+1} - x^*\|_Q^2 = \left(1 - \frac{\nabla f_k^T \nabla f_k}{(\nabla f_k^T Q \nabla f_k)(\nabla f_k^T Q^{-1} \nabla f_k)}\right) \|x_k - x^*\|_Q^2$$

$$x_{k+1} = x_k - \frac{\nabla_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k} \nabla f_k = x_k - \alpha_k \nabla f_k \qquad (3.26)$$

$$\implies x_k - x_{k+1} = \alpha_k \nabla f_k$$

Also $s^T Q y = (s^T Q y)^T = y^T Q^T (s^T)^T = y^T Q s$ because $s^T Q y$ is a number (transpose no effect) and $Q$ is SPD (symmetric)

$$\begin{aligned}
\|x_k - x^*\|_Q^2 - \|x_{k+1} - x^*\|_Q^2 &= (x_k - x^*)^T Q (x_k - x^*) - (x_{k+1} - x^*)^T Q (x_{k+1} - x^*) && \text{by def of } \|.\|_Q \\
&= x_k^T Q x_k - 2x_k^T Q x^* + x^{*T} Q x^* - (x_{k+1}^T Q x_{k+1} - 2x_{k+1}^T Q x^* + x^{*T} Q x^*) && s^T Q y = y^T Q s \\
&= x_k^T Q x_k - 2x_k^T Q x^* - x_{k+1}^T Q x_{k+1} + 2x_{k+1}^T Q x^* \\
&= x_k^T Q x_k - x_{k+1}^T Q x_{k+1} - 2x_k^T b + 2x_{k+1}^T b && Qx^* = b \text{ unique minimizer} \\
&= (x_k - x_{k+1})^T Q (x_k + x_{k+1}) - 2b^T (x_k - x_{k+1}) && s^T t = t^T s \text{ for vectors} \\
&= (\alpha_k \nabla f_k)^T Q (\alpha_k \nabla f_k + 2x_{k+1}) - 2b^T (\alpha_k \nabla f_k) && \text{by 3.26} \\
&= \alpha_k^2 \nabla f_k^T Q \nabla f_k + 2\alpha_k \nabla f_k^T Q x_{k+1} - 2\alpha_k \nabla f_k^T b && \text{can pull out scalars} \\
&= \alpha_k^2 \nabla f_k^T Q \nabla f_k + 2\alpha_k \nabla f_k^T (Q x_{k+1} - b) \\
&= \alpha_k^2 \nabla f_k^T Q \nabla f_k + 2\alpha_k \nabla f_k^T (Q x_{k+1} - Q x^*) && Qx^* = b \\
&= \alpha_k^2 \nabla f_k^T Q \nabla f_k + 2\alpha_k \nabla f_k^T Q (x_{k+1} - x^*) \\
&= \alpha_k^2 \nabla f_k^T Q \nabla f_k + 2\alpha_k \nabla f_k^T Q (x_k - \alpha_k \nabla f_k - x^*) && \text{by 3.26} \\
&= \alpha_k^2 \nabla f_k^T Q \nabla f_k + 2\alpha_k \nabla f_k^T Q (x_k - x^*) - 2\alpha_k^2 \nabla f_k^T Q \nabla f_k \\
&= 2\alpha_k \nabla f_k^T Q (x_k - x^*) - \alpha_k^2 \nabla f_k^T Q \nabla f_k
\end{aligned}$$

Will use $\nabla f_k = Q x_k - b = Q x_k - Q x^* = Q(x_k - x^*) \qquad \implies (x_k - x^*) = Q^{-1} \nabla f_k \qquad$ (i)

$$\begin{aligned}
\|x_k - x^*\|_Q^2 - \|x_{k+1} - x^*\|_Q^2 &= 2\alpha_k \nabla f_k^T Q (x_k - x^*) - \alpha_k^2 \nabla f_k^T Q \nabla f_k \\
&= 2 \frac{\nabla_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k} \nabla f_k^T \nabla f_k - \left(\frac{\nabla_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k}\right)^2 \nabla f_k^T Q \nabla f_k && \text{by def of } \alpha_k \text{ 3.26 and (i)} \\
&= \frac{2(\nabla_k^T \nabla f_k)^2}{\nabla f_k^T Q \nabla f_k} - \frac{(\nabla_k^T \nabla f_k)^2}{\nabla f_k^T Q \nabla f_k} && \nabla f_k^T Q \nabla f_k \text{ is a scalar} \\
&= \frac{(\nabla_k^T \nabla f_k)^2}{\nabla f_k^T Q \nabla f_k}
\end{aligned}$$

$$\begin{aligned}
\|x_k - x^*\|_Q^2 &= (x_k - x^*)^T Q (x_k - x^*) && \text{by def of } \|.\|_Q \\
&= (x_k - x^*)^T \nabla f_k && \text{by (i)} \\
&= (Q^{-1} \nabla f_k)^T \nabla f_k && \text{by (i)} \\
&= \nabla f_k^T (Q^{-1})^T \nabla f_k && \text{by prop of } T \\
&= \nabla f_k^T Q^{-1} \nabla f_k && Q^{-1} \text{ is symmetric since } Q \text{ is}
\end{aligned}$$

Finally, by the last two qualities

$$\begin{aligned}\left\|x_{k+1}-x^*\right\|_Q^2 &= \left\|x_k-x^*\right\|_Q^2 - \frac{(\nabla_k^T\nabla f_k)^2}{\nabla f_k^T Q\nabla f_k}\\ &= \left(1-\frac{(\nabla_k^T\nabla f_k)^2}{\nabla f_k^T Q\nabla f_k\|x_k-x^*\|_Q^2}\right)\left\|x_k-x^*\right\|_Q^2\\ &= \left(1-\frac{(\nabla_k^T\nabla f_k)^2}{\nabla f_k^T Q\nabla f_k\nabla f_k^T Q^{-1}\nabla f_k}\right)\left\|x_k-x^*\right\|_Q^2\end{aligned}$$

Using Kantorovitch Inequality, when Q is SPD, we have for all x

$$\frac{(x^Tx)^2}{(x^TQx)(x^TQ^{-1}x)} \geq \frac{4\lambda_n\lambda_1}{(\lambda_n+\lambda_1)^2}$$

$$\begin{aligned}\left\|x_{k+1}-x^*\right\|_Q^2 &= \left(1-\frac{(\nabla_k^T\nabla f_k)^2}{\nabla f_k^T Q\nabla f_k\nabla f_k^T Q^{-1}\nabla f_k}\right)\left\|x_k-x^*\right\|_Q^2\\ &\leq \left(1-\frac{4\lambda_n\lambda_1}{(\lambda_n+\lambda_1)^2}\right)\left\|x_k-x^*\right\|_Q^2\\ &= \left(\frac{(\lambda_n+\lambda_1)^2-4\lambda_n\lambda_1}{(\lambda_n+\lambda_1)^2}\right)\left\|x_k-x^*\right\|_Q^2\\ &= \left(\frac{\lambda_n^2+2\lambda_n\lambda_1+\lambda_1^2-4\lambda_n\lambda_1}{(\lambda_n+\lambda_1)^2}\right)\left\|x_k-x^*\right\|_Q^2\\ &= \left(\frac{\lambda_n^2-2\lambda_n\lambda_1+\lambda_1^2}{(\lambda_n+\lambda_1)^2}\right)\left\|x_k-x^*\right\|_Q^2\\ &= \left(\frac{(\lambda_n-\lambda_1)^2}{(\lambda_n+\lambda_1)^2}\right)\left\|x_k-x^*\right\|_Q^2\\ &= \left(\frac{(\lambda_n-\lambda_1)(\lambda_n+\lambda_1)}{(\lambda_n+\lambda_1)^2}\right)\left\|x_k-x^*\right\|_Q^2\\ &= \left(\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}\right)\left\|x_k-x^*\right\|_Q^2\end{aligned}$$

# 3 Question 3

Consider the problem of minimizing $f(x) = e^x$ for $x \in \mathbf{R}$

a) Consider the line-search steepest-descent method with $x_0 = 0$ and $\alpha_k = 1$ for all $k = 0, 1, 2...$
First show that, with this choice of $\alpha_k$, $x_k \to \infty$ as $k \to \infty$.
Next show that, with this choice of $\alpha_k$, the line-search steepest-descent method does not satisfy the Wolfe conditions (3.6a)–(3.6b) on page 34 of the Nocedal and Wright textbook for any choice of the constants $c_1$ and $c_2$ satisfying $0 < c_1 < c_2 < 1$.

Steepest descent method forms the iterates $x_k$ by

$$x_{k+1} = x_k + \alpha_k p_k \qquad \text{where } p_k = -\nabla f(x_k)$$

Here $\nabla f(x_k) = f'(x_k) = -e^{-x_k} = -f(x_k)$ and $\alpha_k = 1$
So $x_{k+1} = x_k + f(x_k) = x_k + e^{-x_k}$
Clearly all the $x_k \geq 0$ and the $x_k$ keep increasing.
(Base case: $x_0 = 0 \geq 0$, Induction step: if $x_k >= 0$ then $x_{k+1} = x_k + e^{-x_k} > x_k \geq 0$ since $e^{-x_k} = f(x_k) > 0$ for all x)
So $\lim_{k\to\infty} x_k$ must either be inifinity or a finite number, can't oscillate around a number since they are only increasing).

Define

$$P(k) : x_k = \sum_{i=0}^{k-1} e^{-x_i}$$

Will prove that $P(k)$ holds for all k
Base Case: Let $k = 0$, $x_k = x_0 = 0$ given. $\sum_{i=0}^{k-1} e^{-x_k} = 0$, empty sum. Therefore $P(0)$ holds
Induction Step: Let $k \in \mathbf{N}$ and assume that $P(k)$ holds. That is $x_k = \sum_{i=0}^{k-1} e^{-x_i}$
WTP $P(k+1)$ holds, that is $x_{k+1} = \sum_{i=0}^{k} e^{-x_i}$

$$x_{k+1} = x_k + e^{-x_k} = \sum_{i=0}^{k-1} e^{-x_i} + e^{-x_k} = \sum_{i=0}^{k} e^{-x_i} \qquad \text{by def of } x_{k+1} \text{ and I.H.}$$

Hence $P(k+1)$ holds, and therefore for all k, $x_k = \sum_{i=0}^{k-1} e^{-x_i}$

Now assume that the $x_k$ are bounded, that is for all k $x_k < L$ for some finite number L
Since $x_k < L$ and $e^{-x}$ is a strictly decreasing function, we have that $e^{-x_k} > e^{-L}$ for all k
Now consider the limit ($Le^L$ is a finite number too)

$$\lim_{k\to\infty} x_k = \lim_{k\to\infty} \sum_{i=0}^{k-1} e^{-x_i} > \lim_{k\to\infty} \sum_{i=0}^{k-1} e^{-L} = e^{-L} \lim_{k\to\infty} \sum_{i=0}^{k-1} 1 > e^{-L} \frac{L}{e^{-L}} = L$$

Meaning $x_k$ will eventually be larger than L, which is a contradiction since we assumed all the $x_k$ are bounded above by L. Hence $x_k \to \infty$ as $k \to \infty$

Now to show that the choice $a_k = 1$ does not satisfy Wolfe condition (3.6b) for all choices of $c_1$ and $c_2$

$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f(x_k)^T p_k$$

Here $a_k = 1$, $f'(x) = -f(x)$ and $p_k = -f'(x_k) = f(x_k)$

$$\begin{aligned}
\nabla f(x_k + \alpha_k p_k)^T p_k &\geq c_2 \nabla f(x_k)^T p_k & \\
f'(x_k + \alpha_k p_k) p_k &\geq c_2 f'(x_k) p_k & \text{in 1D} \\
-f(x_k + f(x_k)) f(x_k) &\geq -c_2 f(x_k) f(x_k) & \text{by def of } a_k,\ f'(x) \text{ and } p_k \\
f(x_k + f(x_k)) f(x_k) &\leq c_2 f(x_k) f(x_k) & \text{multiply by } -1 \\
f(x_k + f(x_k)) &\leq c_2 f(x_k) & \text{divide out by } f(x_k), \text{ always positive} \\
e^{-(x_k + e^{-x_k})} &\leq c_2 e^{-x_k} & \text{by def of } f(x) \\
e^{-x_k} e^{-e^{-x_k}} &\leq c_2 e^{-x_k} & \\
e^{-e^{-x_k}} &\leq c_2 & e^{-x} \text{ always positive, no change in sign}
\end{aligned}$$

But before it was show that $x_k \to \infty$, so $e^{-x_k} \to 0$, and so $e^{-e^{-x_k}} \to 1$
But $c_2$ must be $< 1$. No matter how large $c_2 < 1$ is set to be, eventually $e^{-e^{-x_k}}$ will equal $c_2$.
Meaning no matter what we set $c_2$ to be, it will eventually have to be larger than $c_2$, so there is no $c_2$ that will work.

b) Consider the line-search steepest-descent method with $x_0 = 0$ and $\alpha_k = e^{x_k}$ for all $k = 0, 1, 2...$
First show that, with this choice of $\alpha_k$, $x_k \to \infty$ as $k \to \infty$.
Next show that, with this choice of $\alpha_k$, the line-search steepest-descent method does satisfy the strong Wolfe conditions (3.7a)–(3.7b) on page 34 of the Nocedal and Wright textbook for some constants $c_1$ and $c_2$ satisfying $0 < c_1 < c_2 < 1$.
Are there any additional constraints on the constants $c_1$ and $c_2$ other than $0 < c_1 < c_2 < 1$?

Like before $p_k = f(x_k) = e^{-x_k}$
By the steepest descent update, $x_{k+1} = x_k + \alpha_k p_k = x_k + e^{x_k} e^{-x_k} = x_k + 1$
Define $P(k) : x_k = k$
Base Case: $k = 0$, $x_k = x_0 = 0 = k$ by def. So $P(0)$ holds.
Induction Step: Let $k \in \mathbf{N}$ and assume that $P(k)$ holds, that is $x_k = k$
$x_{k+1} = x_k + 1 = k + 1$, so $P(k+1)$ holds.
Hence for all k = 0, 1.., $x_k = k$

$$\lim_{k\to\infty} x_k = \lim_{k\to\infty} k = \infty$$

Next will show that the strong Wolfe conditions hold.

$$
\begin{aligned}
f(x_k+\alpha_k p_k) &\le f(x_k)+c_1\alpha_k\nabla f(x_k)^T p_k && \text{Wolfe condition 3.7a}\\
f(x_k+\alpha_k p_k) &\le f(x_k)+c_1\alpha_k f'(x_k)p_k && \text{condition in 1D}\\
f(x_k+1) &\le f(x_k)+c_1 f'(x_k) && \alpha_k p_k = e^{x_k}e^{-x_k}=1\\
f(x_k+1) &\le f(x_k)-c_1 f(x_k) && f'(x)=-f(x)\\
e^{-x_k-1} &\le e^{-x_k}-c_1 e^{-x_k} && f(x)=e^{-x}\\
e^{-x_k}e^{-1} &\le e^{-x_k}-c_1 e^{-x_k} &&\\
e^{-1} &\le 1-c_1 && f(x)=e^{-x}\text{ always positive}\\
c_1 &\le 1-e^{-1}\approx 0.6321 &&
\end{aligned}
$$

$$
\begin{aligned}
\left|\nabla f(x_k+\alpha_k p_k)^T p_k\right| &\le c_2\left|\nabla f(x_k)^T p_k\right| && \text{Wolfe condition 3.7b}\\
\left|f'(x_k+\alpha_k p_k)p_k\right| &\le c_2\left|f'(x_k)p_k\right| && \text{in 1D}\\
\left|f'(x_k+1)(-f'(x_k))\right| &\le c_2\left|f'(x_k)(-f'(x_k))\right| && \text{by def of }\alpha_k\text{ and }p_k\\
\left|-f(x_k+1)f(x_k)\right| &\le c_2\left|-f(x_k)f(x_k)\right| && f'(x)=-f(x)\\
f(x_k+1)f(x_k) &\le c_2 f(x_k)f(x_k) && f(x)\text{ always positive}\\
f(x_k+1) &\le c_2 f(x_k) && f(x)\text{ always positive, no sign change}\\
e^{-(x_k+1)} &\le c_2 e^{-x_k} && \text{by def of } f(x)\\
e^{-x_k}e^{-1} &\le c_2 e^{-x_k} &&\\
0.3679\approx e^{-1} &\le c_2 &&
\end{aligned}
$$

This allows for a range of acceptable $c_1$ and $c_2$, with the two additional constraints on them above.

c) Consider the line-search Newton's method with $x_0=0$ and $\alpha_k=1$ for all $k=0,1,2...$
First show that, with this choice of $\alpha_k$, $x_k\to\infty$ as $k\to\infty$.
Next show that, with this choice of $\alpha_k$, the line-search Newton's method does satisfy the strong Wolfe conditions (3.7a)–(3.7b) on page 34 of the Nocedal and Wright textbook for some constants $c_1$ and $c_2$ satisfying $0<c_1<c_2<1$.
Are there any additional constraints on the constants $c_1$ and $c_2$ other than $0<c_1<c_2<1$?

$f(x)=e^{-x}$, $f'(x)=-e^{-x}$ and $f''(x)=e^{-x}=f(x)$
By Newton method, $x_{k+1}=x_k+\alpha_k p_k^N$
where $p_k^N$ satisfies $-\nabla^2 f(x_k)p_k^N=\nabla f(x_k)\implies -f''(x_k)p_k^N=f'(x_k)\implies -f(x_k)p_k^N=-f(x_k)\implies p_k^N=1$
So $x_{k+1}=x_k+1$
Like was shown in b), this $x_k$ satisfies $x_k=k$ and so

$$\lim_{k\to\infty}x_k=\lim_{k\to\infty}k=\infty$$

$$
\begin{aligned}
f(x_k+\alpha_k p_k) &\le f(x_k)+c_1\alpha_k\nabla f(x_k)^T p_k && \text{Wolfe condition 3.7a}\\
f(x_k+\alpha_k p_k) &\le f(x_k)+c_1\alpha_k f'(x_k)p_k && \text{condition in 1D}
\end{aligned}
$$

Is satisfied the same way as in b), since $\alpha_k p_k=1$ here as well. Gives $c_1\le 1-e^{-1}\approx 0.6321$

$$
\begin{aligned}
\left|\nabla f(x_k+\alpha_k p_k)^T p_k\right| &\le c_2\left|\nabla f(x_k)^T p_k\right| && \text{Wolfe condition 3.7b}\\
\left|f'(x_k+\alpha_k p_k)p_k\right| &\le c_2\left|f'(x_k)p_k\right| && \text{in 1D}\\
\left|f'(x_k+1)\right| &\le c_2\left|f'(x_k)\right| && \text{by def of }\alpha_k\text{ and }p_k\\
\left|-f(x_k+1)\right| &\le c_2\left|-f(x_k)\right| && f'(x)=-f(x)\\
f(x_k+1) &\le c_2 f(x_k) && f(x)\text{ always positive}\\
e^{-(x_k+1)} &\le c_2 e^{-x_k} && \text{by def of } f(x)\\
e^{-x_k}e^{-1} &\le c_2 e^{-x_k} &&\\
0.3679\approx e^{-1} &\le c_2 &&
\end{aligned}
$$

Clearly there are $c_1$ and $c_2$ that satisfy the same additional constraints as in b).