backtrack-tracking line search

Newton's Method for Unconstrained Problems:
Use Newton's algorithm with backtrack-tracking line search to minimize the function $f(x_1; x_2) = \exp(x_1 + 2x_2) + \exp(x_1 - 2x_2) + \exp(-x_1)$. Use backtracking line search.
Show a plot of the function value minus p^* versus the number of iterations.

Analysis of the problem:

$$f(x_1, x_2) = e^{x_1+2x_2} + e^{x_1-2x_2} + e^{-x_1}$$

For future uses, we compute the gradient and hessian of $f(x_1, x_2)$ first:

Gradient: $\nabla f(x_1, x_2) = \begin{bmatrix} e^{x_1+2x_2} + e^{x_1-2x_2} - e^{-x_1} \\ 2e^{x_1+2x_2} - 2e^{x_1-2x_2} \end{bmatrix}$

Hessian: $\nabla^2 f(x_1, x_2) = \begin{bmatrix} e^{x_1+2x_2} + e^{x_1-2x_2} + e^{-x_1} & 2e^{x_1+2x_2} - 2e^{x_1-2x_2} \\ 2e^{x_1+2x_2} - 2e^{x_1-2x_2} & 4e^{x_1+2x_2} + 4e^{x_1-2x_2} \end{bmatrix}$

Algorithm of Newton's Method:

---

given a starting point $x \in \mathbf{dom}\, f$, tolerance $\epsilon > 0$.
repeat

1. *Compute the Newton step and decrement.*
   $\Delta x_{\mathrm{nt}} := -\nabla^2 f(x)^{-1} \nabla f(x); \quad \lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x).$
2. *Stopping criterion.* **quit** if $\lambda^2/2 \le \epsilon$.
3. *Line search.* Choose step size t by backtracking line search.
4. *Update.* $x := x + t\Delta x_{\mathrm{nt}}$.

Step 1: compute the newton's step:

Since $(x + \Delta_{nt})$ is the minimizer of second-order approximation of $f(x)$ at x, second-order Taylor approximation at point x is:

$$\hat{f}(x + \Delta x_{nt}) = f(x) + \nabla f(x)^T \Delta x_{nt} + \frac{1}{2} \Delta x_{nt}{}^T \nabla^2 f(x) \Delta x_{nt}$$

Take the derivative of $\hat{f}(x + \Delta x_{nt})$ with respect to Δx_{nt}, making it equal to zero, we get:

$$\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

Plug in the gradient and hessian computed above,
We have:

$$\Delta x_{nt} = -\begin{bmatrix} e^{x_1+2x_2} + e^{x_1-2x_2} + e^{-x_1} & 2e^{x_1+2x_2} - 2e^{x_1-2x_2} \\ 2e^{x_1+2x_2} - 2e^{x_1-2x_2} & 4e^{x_1+2x_2} + 4e^{x_1-2x_2} \end{bmatrix}^{-1} * \begin{bmatrix} e^{x_1+2x_2} + e^{x_1-2x_2} - e^{-x_1} \\ 2e^{x_1+2x_2} - 2e^{x_1-2x_2} \end{bmatrix}$$

Step 2: Compute stopping criterion (newton decrement):

For the purpose of determining stopping criterion, we need to quantify the distance between the current function value and optimal value, namely: $f(x) - p^*$

However, we are unable to know the value of p^*, then we consider if the decrement of function value on a certain step $f(x) - \inf_{\Delta x_{nt}} \hat{f}(x + \Delta x_{nt})$ is small enough, we will be able to claim we reach the optimal point.

From the above second-order Taylor approximation, we know that

$$
\begin{aligned}
f(x) - \inf_{\Delta x_{nt}} \hat{f}(x + \Delta x_{nt}) &= f(x) - \inf_{\Delta x_{nt}} \left[f(x) + \nabla f(x)^T \Delta x_{nt} + \frac{1}{2} {\Delta x_{nt}}^T \nabla^2 f(x) \Delta x_{nt} \right] \\
&= f(x) \\
&\quad - \Big[f(x) + \nabla f(x)^T \left(-\nabla^2 f(x)^{-1} \nabla f(x)\right) \\
&\quad + \frac{1}{2} \left(-\nabla^2 f(x)^{-1} \nabla f(x)\right)^T \nabla^2 f(x) \left(-\nabla^2 f(x)^{-1} \nabla f(x)\right) \Big] \\
&= \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \\
&\quad - \frac{1}{2} \left(\nabla f(x)\right)^T (\nabla^2 f(x)^{-1})^T \nabla^2 f(x) (\nabla^2 f(x)^{-1}) \left(\nabla f(x)\right) \\
&= \frac{1}{2} \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \overset{\text{def}}{=} \frac{1}{2} \lambda(x)^2
\end{aligned}
$$

$\lambda(x) = [\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)]^{\frac{1}{2}}$ is called newton decrement. Since $\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$, $\lambda(x)$ can also be expressed as: $\lambda(x) = [(\nabla^2 f(x) \Delta x_{nt})^T \nabla^2 f(x)^{-1} (\nabla^2 f(x) \Delta x_{nt})]^{\frac{1}{2}} = [\Delta x_{nt} \nabla^2 f(x) \Delta x_{nt}]^{\frac{1}{2}}$

Δx_{nt} is steepest descent direction at x in local Hessian norm:

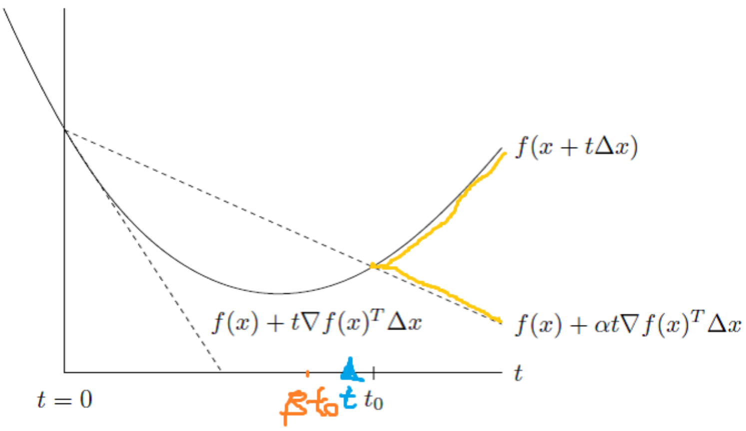
Hessian norm : $\Delta x_{nt} = \|u\|_{\nabla^2 f(x)} = (u^T \nabla^2 f(x) u)^{\frac{1}{2}}$, here $u = \Delta x_{nt}$

Hence, $\lambda(x)$ is the norm (defined by Hessian) of newton step, i.e.

$$\lambda(x) = \|\Delta x_{nt}\|_{\nabla^2 f(x)} = ({\Delta x_{nt}}^T \nabla^2 f(x) \Delta x_{nt})^{\frac{1}{2}}$$

Step 3: compute t through backtracking line search:

For backtracking line search (with $\alpha \in \left(0, \frac{1}{2}\right), \beta \in (0,1)$)



Since $f(x)$ is convex, the linear approximation $f(x) + \alpha t \nabla f(x)^T \Delta x$ must be an

under-estimator of $f(x+t\Delta x)$, i.e. $f(x+t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$
Moreover, as we are doing decrement method, then we have:

$$f(x) > f(x+t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$$

$\nabla f(x)^T \Delta x < 0$, as we already above. The inequality above is to limit the decrement of the current step $t\Delta x$ to be moderate, not too large.

The parameter α is typically chosen between 0.01 and 0.3, meaning that we accept a decrease in f between 1% and 30% of the prediction based on the linear extrapolation $t\nabla f(x)^T \Delta x$.

Why we have both α and t to adjust the amount of decrement? Are they redundant? No! Because α is a pre-determined parameter which is invariant in different steps, while t is a dynamic parameter, which changes as the decrement process goes on.

The problem boils down to: choosing t (at each step):
We want to choose a good t, so that $f(x) + \alpha t \nabla f(x)^T \Delta x$ is as close as to $f(x + t\Delta x)$, i.e. we are to get the smallest t possible (under the premise of satisfying $f(x+t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$)

- If we start from t=1, which satisfying the inequality $f(x+t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$, then we repeat $t := \beta t$ until: $f(x+t\Delta x) \leq f(x) + \alpha t \nabla f(x)^T \Delta x$
- If the starting point t=1 has already broken the inequality, we stop at t=1

Then we get the "good" t, denoted (blue triangle) in the picture above. Since the t is changing discontinuously, the "good" t may not accurately lie on t_0, it will be the first step that is smaller than t_0, which is $(\beta t_0, t_0]$

Step 4: update $\boldsymbol{x} = \boldsymbol{x} + \Delta \boldsymbol{x}_{nt}$