Problem 1 (graded by Yiran) - 50 points

(a) 4 points

In a class, among 20 students, 8 are female, and 12 are male. 2 of the female students are taller than 170 cm, and 8 of the male students are taller than 170 cm. Suppose we randomly pick a student, let

x: the student is female;

y: the student is taller than 170 cm.

Then,

$P(x,y)$ is the probability that the student is both female and is taller than 170 cm, which is equal to $2/20 = 0.1$.

$P(x)$ is the probability that the student is female, which is equal to $8/20 = 0.4$.

$P(y|x)$ is the probability that the student is taller than 170 cm, given it is known the student is female, which is equal to $2/8 = 0.25$.

$P(y)$ is the probability that the student is taller than 170 cm, which is equal to $(2+8)/20 = 0.5$.

$P(x|y)$ is the probability that the student is female, given it is known the student is taller than 170 cm, which is equal to $2/(2+8) = 0.2$.

We see that:

$P(x,y) = P(y|x)P(x) = P(x|y)P(y)$

(b) 8 points

Independent

Let

x : I get an A for Ge/ESE118.

y: The next president of the U.S. is Republican.

These two events are independent, because if x happens, does not affect the probability of y, and vice versa.

Let's assume $P(x) = 3/5$, and $P(y) = 1/2$.

Suppose I get an A with $P(x)$. It doesn't affect the election at all, and there is still 1 in 2 odds that the next president will be Republican. Therefore, to make both happen, $P(x,y) = P(x)P(y)$. Similaly, suppose the Republican wins the election with $P(y)$. It doesn't affect my odd to get an A, and to make both happen, $P(x,y) = P(y)P(x)$.

Intuitively, the rule holds because the two events are independent - one happening does not affect the other; therefore, to make both happen, we need to multiply $P(x)$ and $P(y)$.

Dependent

Let

x: The next president of the U.S. is Democratic.

y: The next president of the U.S. is Republican.

These two events are not independent, because if either of them happens, it will affect the proability of the other.

Let's assume $P(x) = P(y) = \frac{1}{2}$. Because it's impossible that the next president is both Democratic and Republican, $P(x,y) = 0 \neq P(x)P(y)$.

(c) 8 points

(c.i) 3 points

$$E(x) = \int_{-\infty}^{\infty} xP(x)dx$$

Since x is an odd function, and $P(x)$ is an even function, their product is an odd function. Integration of an odd function over symmetric boundaries as $[-\infty, \infty]$ is 0. Therefore,

$$E(x) = 0$$

(c.ii) 5 points

$$\begin{aligned} E(x^2) &= \int_{-\infty}^{\infty} x^2 P(x)dx \\ &= \int_{-\infty}^{\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \end{aligned}$$

Since

$$\left[\exp\left(-\frac{x^2}{2\sigma^2}\right)\right]' = -\frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)$$

Then

$$\begin{aligned} E(x^2) &= -\frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \left[\exp\left(-\frac{x^2}{2\sigma^2}\right)\right]' dx \\ &= -\frac{\sigma}{\sqrt{2\pi}} \left[x \exp\left(-\frac{x^2}{2\sigma^2}\right)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \end{aligned}$$

where we use

$$\begin{aligned} \lim_{x\to\infty} x \exp\left(-\tfrac{x^2}{2\sigma^2}\right) &= \lim_{x\to\infty} \frac{x}{\exp\left(\frac{x^2}{2\sigma^2}\right)} \\ &= \lim_{x\to\infty} \frac{x'}{\left[\exp\left(\frac{x^2}{2\sigma^2}\right)\right]'} \\ &= \lim_{x\to\infty} \frac{1}{\exp(\frac{x^2}{2\sigma^2})\frac{x}{\sigma^2}} \\ &= 0 \end{aligned}$$

Since

$$\begin{aligned}
\int_{-\infty}^{\infty} e^{-x^2/2} dx &= \sqrt{2\pi} \\
&= \int_{-\infty}^{\infty} \exp\left[-\frac{\left(\frac{x}{\sigma}\right)^2}{2}\right] d\left(\frac{x}{\sigma}\right) \\
&= \frac{1}{\sigma}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx
\end{aligned}$$

Then

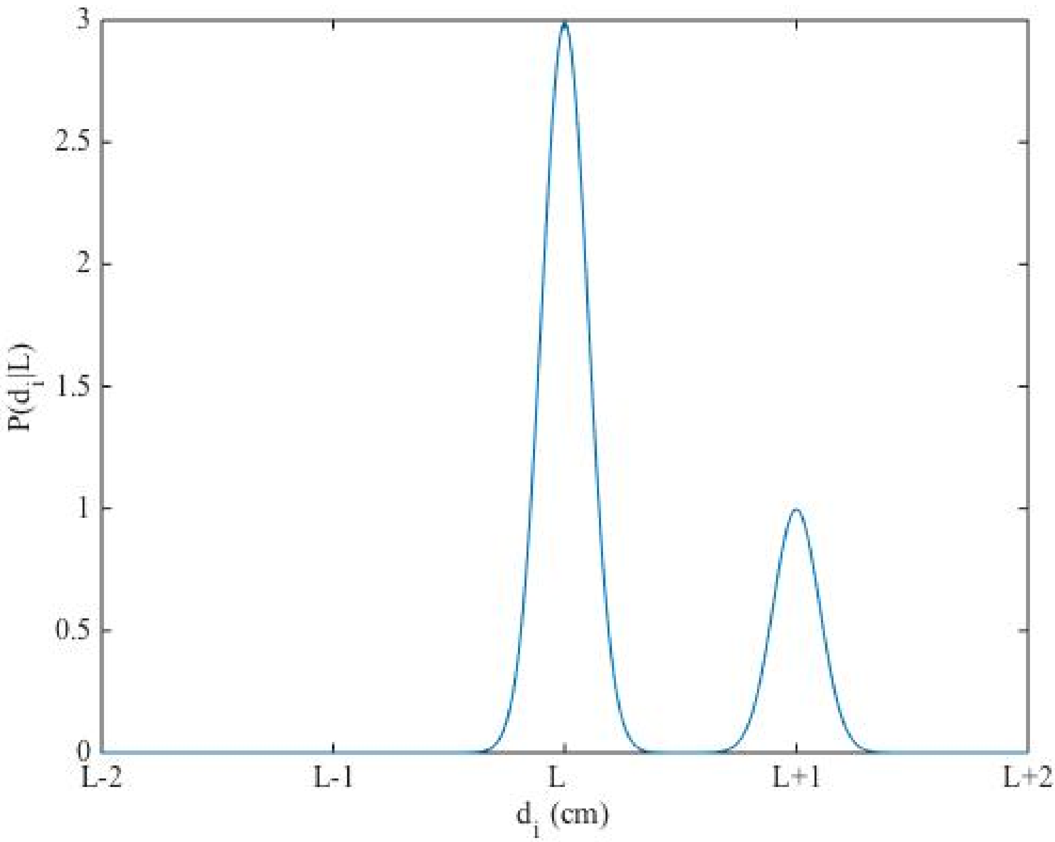
$$E(x^2) = \frac{\sigma}{\sqrt{2\pi}}\sqrt{2\pi}\sigma = \sigma^2$$

(d) 30 points

(d.i) 10 points

$$\begin{aligned}
P(d_i|L) &= \frac{3}{4}P_{Gauss}(d_i|L,\sigma) + \frac{1}{4}P_{Gauss}(d_i|L+1,\sigma) \\
&= \frac{3}{4}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(d_i-L)^2}{2\sigma^2}\right) + \frac{1}{4}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(d_i-(L+1))^2}{2\sigma^2}\right) \\
&= \frac{15}{2}\frac{1}{\sqrt{2\pi}}\exp\left(-50(d_i-L)^2\right) + \frac{5}{2}\frac{1}{\sqrt{2\pi}}\exp\left(-50(d_i-(L+1))^2\right)
\end{aligned}$$

where $\sigma = 0.1$ cm, and L, d_i are in cm.



(d.ii) 10 points

From Bayes' theorem,

$$\begin{aligned}P(L|d_i) &\propto P(d_i|L)P(L)\\ &\propto P(d_i|L) = \frac{3}{4}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(d_i-L)^2}{2\sigma^2}\right)+\frac{1}{4}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(d_i-(L+1))^2}{2\sigma^2}\right)\end{aligned}$$

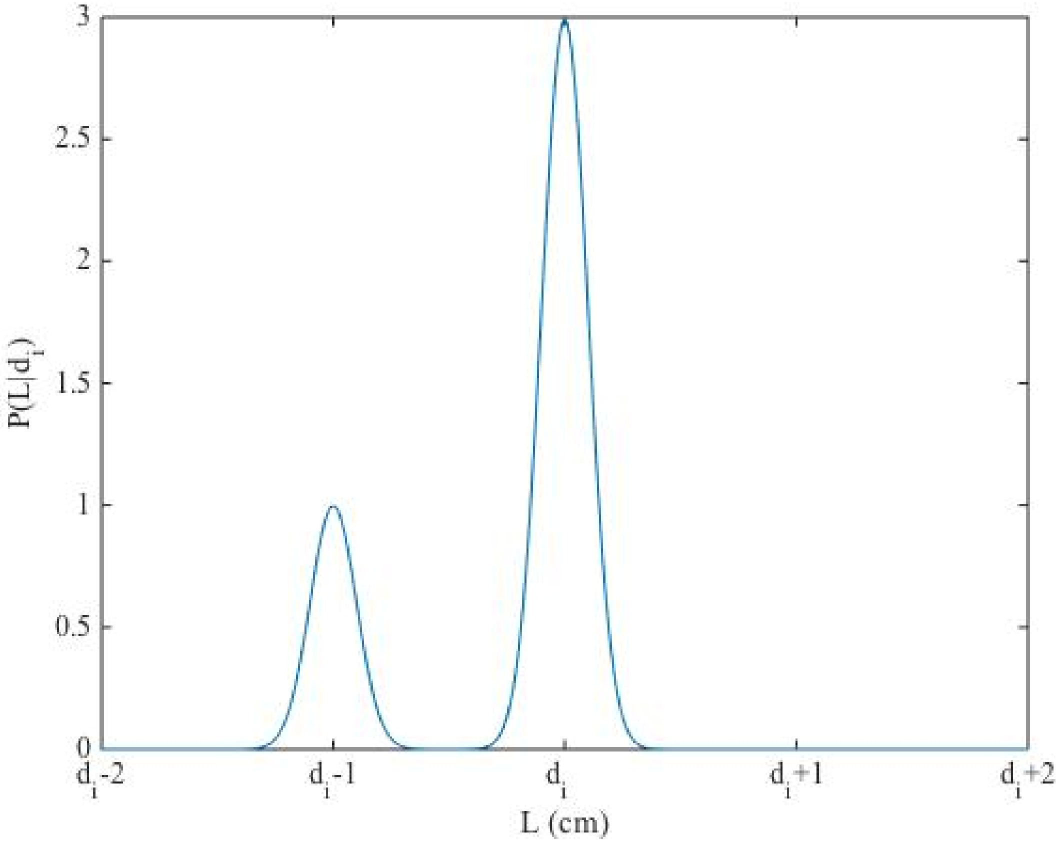
where we assume the prior distribution $P(L)$ is uniform.
Since

$$\begin{aligned}P(d_i|L) &= \frac{3}{4}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(d_i-L)^2}{2\sigma^2}\right)+\frac{1}{4}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(d_i-(L+1))^2}{2\sigma^2}\right)\\ &= \frac{3}{4}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(L-d_i)^2}{2\sigma^2}\right)+\frac{1}{4}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(L-(d_i-1))^2}{2\sigma^2}\right)\\ &= \frac{3}{4}P_{Gauss}(L|d_i,\sigma)+\frac{1}{4}P_{Gauss}(L|d_i-1,\sigma)\end{aligned}$$

Then

$$\int_{-\infty}^{\infty} P(d_i|L)dL \quad = 1$$

$$
\begin{aligned}
P(L|d_i) &= P(d_i|L) \\
&= \frac{3}{4}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(L-d_i)^2}{2\sigma^2}\right)+\frac{1}{4}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(L-(d_i-1))^2}{2\sigma^2}\right) \\
&= \frac{15}{2}\frac{1}{\sqrt{2\pi}}\exp\left(-50(L-d_i)^2\right)+\frac{5}{2}\frac{1}{\sqrt{2\pi}}\exp\left(-50(L-(d_i-1))^2\right)
\end{aligned}
$$



(d.iii) 10 points

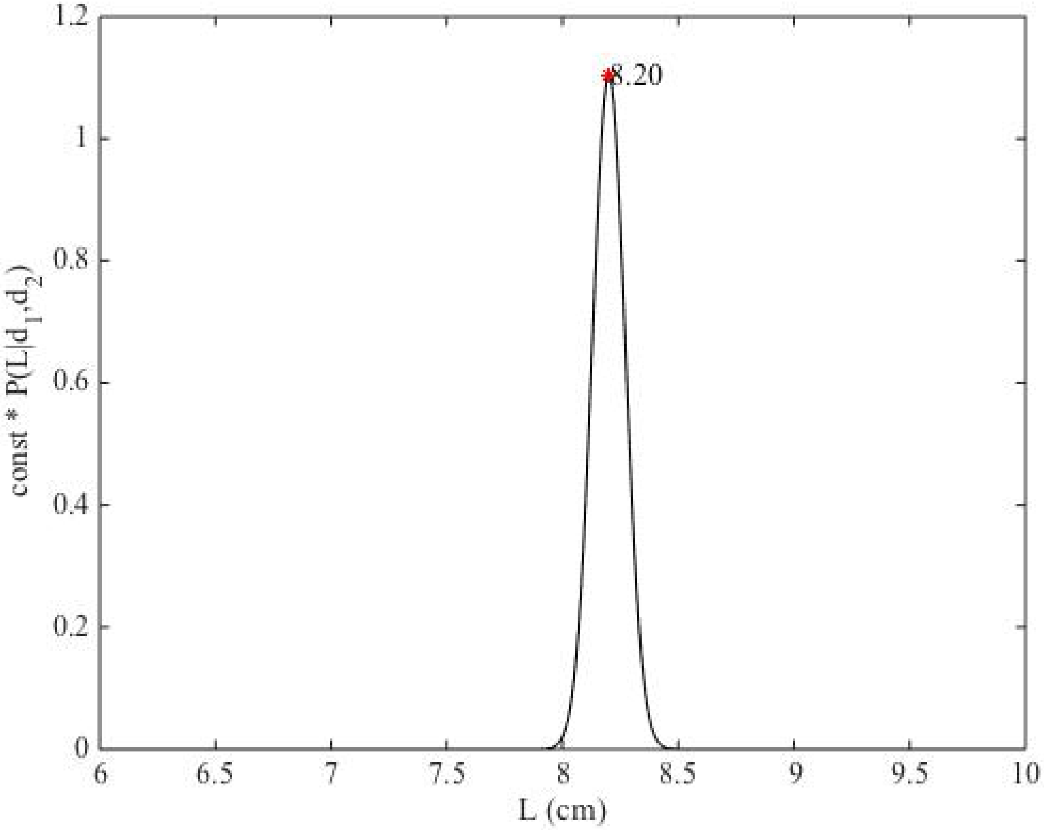
From Bayes' theorem,

$$
\begin{aligned}
P(L|\{d_1=8.3,d_2=9.1\}) &\propto P(\{d_1=8.3,d_2=9.1\}|L)P(L) \\
&\propto P(\{d_1=8.3,d_2=9.1\}|L) \\
&= P(d_1|L)P(d_2|L)
\end{aligned}
$$

Since we only care about the maximum of the LHS, instead of its value; for the RHS, absorbing the $\frac{1}{4}\frac{1}{\sigma\sqrt{2\pi}}$ terms into the constant

$$
\begin{aligned}
P(L|\{d_1=8.3,d_2=9.1\}) \propto\ & \left[3\exp\left(-50(8.3-L)^2\right)+\exp\left(-50(7.3-L)^2\right)\right] \\
& \cdot\left[3\exp\left(-50(9.1-L)^2\right)+\exp\left(-50(8.1-L))^2\right)\right]
\end{aligned}
$$

Use MATLAB to plot the function, and find its maximum for $L = [6 : 0.01 : 10]$. We see that the best estimate of L is 8.2 cm.



Because the two measurements differ ≈ 1, it's likely that in the second measurement, the one quarter chance of additional 1 cm happens. After the 1 cm correction, the second measurement should be 8.1. The mean of 8.3 and 8.1 is 8.2, which is our estimation through the analysis above.

Problem 2 (graded by Kangchen) - 50 points

(a)12 points

According to Bayes' Theorem:

$$P(\boldsymbol{m}|\mathbf{d}) \propto P(\boldsymbol{d}|\boldsymbol{m})P(\boldsymbol{m})$$

we assume a uniform prior distribution $P(\boldsymbol{m})$ equals constant.

$$P(\boldsymbol{m}|\boldsymbol{d}) \propto P(\boldsymbol{d}|\boldsymbol{m})P(\boldsymbol{m})$$

$$P(\boldsymbol{m}|\{d_1, d_2, ..., d_n\}) \propto P(\{d_1, d_2, ..., d_n\}|\boldsymbol{m}) = P(d_1|\boldsymbol{m})P(d_2|\boldsymbol{m})P(d_3|\boldsymbol{m})...P(d_n|\boldsymbol{m})$$

$$P(d_k|\boldsymbol{m}) = e^{-\frac{(d_k - g_k(\boldsymbol{m}))^2}{2\sigma_k^2}}$$

So we multiply these terms together:
$P(\boldsymbol{m}|\boldsymbol{d}) \propto e^{-F(\boldsymbol{m})}$ where $F(\boldsymbol{m}) = \sum \frac{(d_k - g_k(\boldsymbol{m}))^2}{2\sigma_k^2}$

(b)12 points

Since $d'_k = d_k/\sigma_k$, the relation between $\boldsymbol{d'}$ and $\boldsymbol{d}$ can be written in matrix form $\boldsymbol{d'} = \boldsymbol{W}\boldsymbol{d}$ where

$$\boldsymbol{W} = \begin{bmatrix} \frac{1}{\sigma_1} & & & & \\ & \frac{1}{\sigma_2} & & & \\ & & \frac{1}{\sigma_3} & & \\ & & & ... & \\ & & & & \frac{1}{\sigma_k} \end{bmatrix} \text{ and } \boldsymbol{W}\ \boldsymbol{d} = \begin{bmatrix} \frac{1}{\sigma_1} & & & & \\ & \frac{1}{\sigma_2} & & & \\ & & \frac{1}{\sigma_3} & & \\ & & & ... & \\ & & & & \frac{1}{\sigma_k} \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ ... \\ d_k \end{bmatrix} = \begin{bmatrix} \frac{d_1}{\sigma_1} \\ \frac{d_2}{\sigma_2} \\ \frac{d_3}{\sigma_3} \\ ... \\ \frac{d_k}{\sigma_k} \end{bmatrix}$$

Similarly, we have $\boldsymbol{g'} = \boldsymbol{W}\boldsymbol{g}$ since

$$\boldsymbol{W}\ \boldsymbol{g} = \begin{bmatrix} \frac{1}{\sigma_1} & & & & \\ & \frac{1}{\sigma_2} & & & \\ & & \frac{1}{\sigma_3} & & \\ & & & ... & \\ & & & & \frac{1}{\sigma_k} \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ ... \\ g_k \end{bmatrix} = \begin{bmatrix} \frac{g_1}{\sigma_1} \\ \frac{g_2}{\sigma_2} \\ \frac{g_3}{\sigma_3} \\ ... \\ \frac{g_k}{\sigma_k} \end{bmatrix}$$

(c)12 points

$$F = \frac{1}{2}\sum(d'_k - g'_k(\boldsymbol{m}))^2 = (\boldsymbol{d'} - \boldsymbol{g'}(\boldsymbol{m}))^T(\boldsymbol{d'} - \boldsymbol{g}'(\boldsymbol{m}))$$

the gradient:

$$\boldsymbol{\nabla} F = \hat{\boldsymbol{G}'}^T(\boldsymbol{d'} - \boldsymbol{g}'(\boldsymbol{m}))$$

the approximate Hessian:

$$\boldsymbol{H} = (\hat{\boldsymbol{G}'}^T\hat{\boldsymbol{G}'})$$

So the least squares solution:

$$\Delta \boldsymbol{m} = (\hat{\boldsymbol{G}'}^T \hat{\boldsymbol{G}'})^{-1} \hat{\boldsymbol{G}'}^T (\boldsymbol{d}' - \boldsymbol{g}'(\boldsymbol{m}))$$

$$\boldsymbol{m} = \boldsymbol{m}_0 + \Delta \boldsymbol{m}$$

since

$$\hat{G}'_{jl} = \frac{\partial g'_j}{\partial m_l} = \frac{1}{\sigma_j} \frac{\partial g_j}{\partial m_l} = \frac{1}{\sigma_j} \hat{G}_{jl}$$

$$\begin{bmatrix} \frac{1}{\sigma_1} & & & & \\ & \frac{1}{\sigma_2} & & & \\ & & \frac{1}{\sigma_3} & & \\ & & & \cdots & \\ & & & & \frac{1}{\sigma_k} \end{bmatrix} \begin{bmatrix} \frac{\partial g_1}{\partial m_1} & \frac{\partial g_1}{\partial m_2} & \cdots & \frac{\partial g_1}{\partial m_k} \\ \frac{\partial g_2}{\partial m_1} & \frac{\partial g_2}{\partial m_2} & \cdots & \frac{\partial g_2}{\partial m_k} \\ \frac{\partial g_3}{\partial m_1} & \frac{\partial g_3}{\partial m_2} & \cdots & \frac{\partial g_3}{\partial m_k} \\ \frac{\partial g_4}{\partial m_1} & \frac{\partial g_4}{\partial m_2} & \cdots & \frac{\partial g_4}{\partial m_k} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial g_n}{\partial m_1} & \frac{\partial g_n}{\partial m_2} & \cdots & \frac{\partial g_n}{\partial m_k} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_1}\frac{\partial g_1}{\partial m_1} & \frac{1}{\sigma_1}\frac{\partial g_1}{\partial m_2} & \cdots & \frac{1}{\sigma_1}\frac{\partial g_1}{\partial m_k} \\ \frac{1}{\sigma_2}\frac{\partial g_2}{\partial m_1} & \frac{1}{\sigma_2}\frac{\partial g_2}{\partial m_2} & \cdots & \frac{1}{\sigma_2}\frac{\partial g_2}{\partial m_k} \\ \frac{1}{\sigma_3}\frac{\partial g_3}{\partial m_1} & \frac{1}{\sigma_3}\frac{\partial g_3}{\partial m_2} & \cdots & \frac{1}{\sigma_3}\frac{\partial g_3}{\partial m_k} \\ \frac{1}{\sigma_4}\frac{\partial g_4}{\partial m_1} & \frac{1}{\sigma_4}\frac{\partial g_4}{\partial m_2} & \cdots & \frac{1}{\sigma_4}\frac{\partial g_4}{\partial m_k} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{\sigma_n}\frac{\partial g_n}{\partial m_1} & \frac{1}{\sigma_n}\frac{\partial g_n}{\partial m_2} & \cdots & \frac{1}{\sigma_n}\frac{\partial g_n}{\partial m_k} \end{bmatrix} = \begin{bmatrix} \frac{\partial g'_1}{\partial m_1} & \frac{\partial g'_1}{\partial m_2} & \cdots & \frac{\partial g'_1}{\partial m_k} \\ \frac{\partial g'_2}{\partial m_1} & \frac{\partial g'_2}{\partial m_2} & \cdots & \frac{\partial g'_2}{\partial m_k} \\ \frac{\partial g'_3}{\partial m_1} & \frac{\partial g'_3}{\partial m_2} & \cdots & \frac{\partial g'_3}{\partial m_k} \\ \frac{\partial' g_4}{\partial m_1} & \frac{\partial' g_4}{\partial m_2} & \cdots & \frac{\partial g'_4}{\partial m_k} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial' g_n}{\partial m_1} & \frac{\partial' g_n}{\partial m_2} & \cdots & \frac{\partial' g_n}{\partial m_k} \end{bmatrix}$$

so

$$\hat{\boldsymbol{G}}' = \boldsymbol{W}\hat{\boldsymbol{G}}'$$

Substitute $\hat{\boldsymbol{G}}' = \boldsymbol{W}\hat{\boldsymbol{G}}$, $\boldsymbol{d}' = \boldsymbol{W}\boldsymbol{d}$, $\boldsymbol{g}' = \boldsymbol{W}\boldsymbol{g}$,

$$\Delta \boldsymbol{m} = (\hat{\boldsymbol{G}}^T \boldsymbol{W}^T \boldsymbol{W} \hat{\boldsymbol{G}})^{-1} \hat{\boldsymbol{G}}^T \boldsymbol{W}^T \boldsymbol{W} (\boldsymbol{d} - \boldsymbol{g}(\boldsymbol{m}))$$

(d)14points

```
x = [0 11 15 6 -7 3]';
y = [0 0 6 13 10 -7]';
d = [0.103 0.162  0.065  0.036 0.025 0.169]';
M0 = [8 -5 10 30]';  %initial guess
Ms = nonlinear_solver(x,y,d,M0,[1,1,0.2,1,2.5,2.5]);
```

```
function [ Grad, Hess] = compute_gradient_approx_hess( x,y,M, misfit ,weight)

W =diag(weight);
xs = M(1);
ys = M(2);
zs = M(3);
p = M(4);

eta = ((x - xs).^2 + (y - ys).^2 + zs^2);

dx = x-xs;
dy = y-ys;

Ghat(:,1) = (3.*p.*zs.*(dx))./((eta).^(5/2));
Ghat(:,2) = (3.*p.*zs.*(dy))./((eta).^(5/2));
Ghat(:,3) = p./(eta).^(3/2) - (3*p.*zs.^2)./(eta).^(5/2);
Ghat(:,4) = zs./(eta).^(3/2);

Grad = (misfit')*(W')*W*Ghat;



Hess = (Ghat')*(W')*W*Ghat;
%this is the apprximated Hessian;



end
```

```

function [M]=nonlinear_solver(x,y,ui,Minit,w)

M=Minit;
misfit = 0;
misfit_old = 0;

for ii = 1:1:1000
```

```
misfit_old = misfit;
misfit=compute_misfit(x,y,M,ui);


[Grad,Hess]=compute_gradient_approx_hess(x,y,M,misfit,w);

deltaM= (Hess)\Grad';

M=M+deltaM;

if ((misfit-misfit_old)'*(misfit-misfit_old)<1e-7)
    break;
end
end
disp(misfit)
disp(ii)
end
```

```
function [ misfit ] = compute_misfit( x,y,M,ui )

xs = M(1);
ys = M(2);
zs = M(3);
p = M(4);
misfit =ui - p*zs./((x - xs).^2 + (y - ys).^2 + zs^2).^(3/2);
end
```

$$\boldsymbol{m} = [8.3068, -5.3425, 11.8179, 31.8569]^T$$

$$\boldsymbol{error} = [-4.99 \times 10^{-5}, 1.20 \times 10^{-5} - 2.95 \times 10^{-3}, 3.59 \times 10^{-4}, -2.87 \times 10^{-5}, 2.74 \times 10^{-6}]$$

The solution is not very different from the previous one $[8.137, -5.142, 11.507, 30.346]^T$. We can find that since we put a smaller weight on station 3, its error is the largest.

Since we put a larger weight on station 5 6 , their errors are smaller.

(Extra Credit) Problem 3 (graded by Yiran) - 25 points

(a) 5 points

The maximum dimension of $\boldsymbol{G}$ spanned by $\{\boldsymbol{g_i}, i = 1...M\}$ is M. Therefore,

$$dim\,(R(\boldsymbol{G})) \;\leq\; M < N = dim(\mathbb{R}^N)$$

(b) 10 points

Since

$$\begin{aligned} \boldsymbol{H} &= \boldsymbol{G}^T\boldsymbol{G} \\ &= \begin{bmatrix} \boldsymbol{g_1}^T \\ \vdots \\ \boldsymbol{g_M}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{g_1} & ... & \boldsymbol{g_M} \end{bmatrix} \end{aligned}$$

then

$$\boldsymbol{H_{ij}} = \boldsymbol{g_i}^T\boldsymbol{g_j}$$

Note that

$$\boldsymbol{g_i}^T\boldsymbol{g_j} = \boldsymbol{g_i} \cdot \boldsymbol{g_j} = ||\boldsymbol{g_i}||_2||\boldsymbol{g_j}||_2 \cos(\angle(\boldsymbol{g_i}, \boldsymbol{g_j}))$$

We see the diagonal elements of $\boldsymbol{H}$ are the squared lengths of the column vectors of $\boldsymbol{G}$. The off-diagonal elements measure how much the column vectors of $\boldsymbol{G}$ project onto each other, which is related to the angle between the two vectors - if the two vectors are orthogonal to each other, the projection equals zero.

(c) 5 points

Similarly,

$$\boldsymbol{G}^T\boldsymbol{d} = (\boldsymbol{g_1}^T\boldsymbol{d}, ..., \boldsymbol{g_M}^T\boldsymbol{d})^T$$

is a column vector whose elements measure how much $\boldsymbol{d}$ projects onto the different vector $\boldsymbol{g_i}$ in the model space.

(d) 5 points

The given equation can be written as

$$\boldsymbol{G}^T(\boldsymbol{G}\boldsymbol{m}_{LLS}) = \boldsymbol{G}^T\boldsymbol{d}$$

where $\boldsymbol{m}_{LLS}$ denotes the linear least squares solution.

This equation implies that $\boldsymbol{G}\boldsymbol{m}_{LLS}$, which is a vector in the model space, equals the projection of $\boldsymbol{d}$ in the model space. Thus, the least squares solution $\boldsymbol{m}_{LLS}$ is the coordinate of $\boldsymbol{d}$ in the model space.