

## Problem 1 - (graded by Toby)

### (a) - 5 points

The range of  $\mathbf{G}$  is defined as the linear combination of  $\mathbf{G}$ 's column vectors. The dimension of the range of  $\mathbf{G}$  is therefore equal to the number of linearly independent column vectors of  $\mathbf{G}$ . Because  $\mathbf{G}$  is an  $N$  by  $M$  matrix, we have at most  $M$  linearly independent columns and the maximum dimension of the range of  $\mathbf{G}$  is  $M$ . The data space is  $\mathbb{R}^N$  and thus has dimension  $N$ . If  $N > M$ . Hence we have that

$$\dim(\text{range}(\mathbf{G})) \leq M < N. \quad (1)$$

This implies that the range of  $\mathbf{G}$  is a subspace of  $\mathbb{R}^N$  whose dimension is less than  $N$ .

### (b) - 5 points

We have that  $H_{ij} = \mathbf{g}_i^T \mathbf{g}_j$ . The diagonal elements of  $\mathbf{H}$  are the squared lengths of the column vectors of  $\mathbf{G}$ . The off diagonal elements can be nonzero and measure how much the column vectors of  $\mathbf{G}$  project onto each other, i.e., what their orientation is with respect to each other. If  $\mathbf{G}$  is an orthogonal matrix, then  $\mathbf{H}$  is a diagonal matrix. If in addition the column vectors have length 1,  $\mathbf{H}$  is the unity matrix as shown in problem 3 of the first homework set. Because

$$\mathbf{g}_i^T \mathbf{g}_j = \|\mathbf{g}_i\|_2 \|\mathbf{g}_j\|_2 \cos(\angle(\mathbf{g}_i, \mathbf{g}_j)), \quad (2)$$

the dot product involves a measure of how much two vectors project onto each other, measured by the cosine factor. As a matter of fact, this is one way of defining the cosine in the first place.

### (c) - 5 points

Similarly,  $\mathbf{G}^T \mathbf{d} = (\mathbf{g}_1^T \mathbf{d}, \dots, \mathbf{g}_M^T \mathbf{d})^T$  is a column vector whose elements measure how much  $\mathbf{d}$  projects onto the different column vectors  $\mathbf{g}_i$  as

$$\mathbf{g}_i^T \mathbf{d} = \|\mathbf{g}_i\|_2 \|\mathbf{d}\|_2 \cos(\angle(\mathbf{g}_i, \mathbf{d})). \quad (3)$$

### (d) - 5 points

Using standard notation and what we know from parts (b) and (c), the equation as given on the homework set can be written as

$$\begin{pmatrix} \mathbf{g}_1^T (m_1 \mathbf{g}_1 + \dots + m_M \mathbf{g}_M) \\ \vdots \\ \mathbf{g}_M^T (m_1 \mathbf{g}_1 + \dots + m_M \mathbf{g}_M) \end{pmatrix} = \begin{pmatrix} \mathbf{g}_1^T \mathbf{d} \\ \vdots \\ \mathbf{g}_M^T \mathbf{d} \end{pmatrix}. \quad (4)$$

Therefore, the least squares solution to this equation  $\mathbf{m}$  implies a vector  $m_1 \mathbf{g}_1 + \dots + m_M \mathbf{g}_M$  in the range of  $\mathbf{G}$  whose projections onto the column vectors of  $\mathbf{G}$  equal the projects of the data vector onto the column vectors of  $\mathbf{G}$ .

## Problem 2 - (graded by Stephen)

### (a) - 5 points

A silly real life example that demonstrates  $\mathbb{P}(x, y) = \mathbb{P}(x|y) \mathbb{P}(y) = \mathbb{P}(y|x) \mathbb{P}(x)$  can involve peanut butter and jelly sandwiches. If we assume the population is sandwiches made in the United States today, then  $\mathbb{P}(x, y)$  is the probability that a given sandwich we pick will be peanut butter and jelly. We can calculate this probability in two different ways. The first is by multiplying the probability that the sandwich has jelly on it  $\mathbb{P}(y)$  (These sandwiches may have jelly and other ingredients, just jelly, or they may have peanut butter and jelly) by the probability that given the sandwich has jelly on it, that it will also have just peanut butter  $\mathbb{P}(x|y)$ . Conversely, we can begin with the probability that the sandwich has peanut butter  $\mathbb{P}(x)$  and multiply by the probability that given the sandwich has peanut butter that it will also have just jelly  $\mathbb{P}(y|x)$ . The probability given jelly that the sandwich will have peanut butter is NOT the same as given peanut butter that it will also have jelly (there could be many more possible sandwiches that include peanut butter for example).

### (b) - 5 points

Independence of events  $x$  and  $y$  implies the  $\mathbb{P}(x, y) = \mathbb{P}(x) \mathbb{P}(y)$ . An example of this would be two strangers going to a sporting event.  $\mathbb{P}(x, y)$  is the probability that both strangers are at the game even though they do not know each other. If these individuals get to the game independently then we can assume the events  $x$  and  $y$  are independent. In this case we can find  $\mathbb{P}(x, y)$  by simply multiplying  $\mathbb{P}(x)$  by  $\mathbb{P}(y)$ . To think about this intuitively we can think that person  $x$  goes to the games  $1/2$  of the time and person  $y$   $1/3$  of the time. If them going to the games is independent then person  $x$  will be at 50 percent of the games. Person  $y$  only goes to 33.3 percent of the games, and because they are not related, will go to 33.3 percent of the games that person  $x$  is at. Thus to find the total percent of games at which both are present, we multiply  $1/2$  by  $1/3$ .

These events would not be independent if person  $x$  and person  $y$  were friends and attended games together, or if they drove to the games together.

### (c) - 5 points

Expectation for  $x$ :

$$\mathbb{E}[x] = \langle x \rangle = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx \quad (5)$$

The Gaussian is an even function and  $x$  is an odd function. Because we are integrating their product (an odd function) over a symmetric interval, we will get an answer of 0. We can also do this explicitly:

$$\begin{aligned} \mathbb{E}[x] &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{x^2}{2\sigma^2}} dx \\ \mathbb{E}[x] &= \frac{1}{\sigma\sqrt{2\pi}} \left( \int_{-\infty}^0 x e^{-\frac{x^2}{2\sigma^2}} dx + \int_0^{\infty} x e^{-\frac{x^2}{2\sigma^2}} dx \right) \\ \mathbb{E}[x] &= \frac{1}{\sigma\sqrt{2\pi}} \left( -\sigma^2 e^{-\frac{x^2}{2\sigma^2}} \Big|_{-\infty}^0 + -\sigma^2 e^{-\frac{x^2}{2\sigma^2}} \Big|_0^{\infty} \right) \\ \mathbb{E}[x] &= \frac{1}{\sigma\sqrt{2\pi}} (-\sigma^2 - 0 + 0 + \sigma^2) \\ \mathbb{E}[x] &= \frac{1}{\sigma\sqrt{2\pi}} (-\sigma^2 + \sigma^2) \\ \mathbb{E}[x] &= 0 \end{aligned}$$

Expectation for  $x^2$ :

$$\mathbb{E}[x^2] = \langle x^2 \rangle = \int_{-\infty}^{\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx \quad (6)$$

This time the Gaussian is multiplied by an even function, so it will not integrate to 0.

$$\mathbb{E}[x] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{x^2}{2\sigma^2}} dx$$

We can do this integral by parts ( $\int u dv = uv - \int v du$ ):

$$u = x \quad du = 1 \\ dv = x e^{-x^2/2\sigma^2} \quad v = -\sigma^2 e^{-x^2/2\sigma^2}$$

$$\mathbb{E}[x] = \frac{1}{\sigma\sqrt{2\pi}} \left( -\sigma x e^{-\frac{x^2}{2\sigma^2}} \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -\sigma^2 e^{-\frac{x^2}{2\sigma^2}} dx \right)$$

The first term will go to zero because the exponential goes to 0 faster than x increases.

$$\mathbb{E}[x] = \frac{1}{\sigma\sqrt{2\pi}} \sigma^2 \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} dx$$

We can now do a variable substitution  $\beta = \frac{x}{\sigma}$ ,  $d\beta = \frac{1}{\sigma} dx$ ,  $dx = \sigma d\beta$  to do this integral. We are now left with the integral given in the problem:

$$\int_{-\infty}^{\infty} \sigma e^{-\frac{\beta^2}{2}} d\beta = \sigma\sqrt{2\pi}$$

We are now left with:

$$\mathbb{E}[x] = \frac{1}{\sigma\sqrt{2\pi}} \sigma^3 \sqrt{2\pi} = \sigma^2$$

ALTERNATE - An alternative way to do this is by using a sneaky derivative trick:

$$\mathbb{E}[x^2] = \langle x^2 \rangle = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2\sigma^2}} dx$$

This looks suspiciously like the derivative of a function with respect to  $\frac{1}{2\sigma^2}$ . Thus, we can do a substitution of  $\alpha = \frac{1}{2\sigma^2}$  and rewrite our integral like so:

$$-\frac{d}{d\alpha} \left( \int_{-\infty}^{\infty} e^{-\alpha x^2} dx \right)$$

We can pull out this derivative because  $\alpha$  has nothing to do with the integral over x. (I agree that this is very sneaky). We already know the answer to the integral on the inside, so we use that and then take the derivative of the answer with respect to  $\alpha$ .

$$-\frac{d}{d\alpha} \left( \sqrt{\frac{\pi}{\alpha}} \right) = \frac{1}{2\alpha} \sqrt{\frac{\pi}{\alpha}}$$

Now plug back in for  $\alpha$  and simplify (don't forget the prefactor we pulled out of the integral at the start).

$$\mathbb{E}[x^2] = \frac{1}{\sigma\sqrt{2\pi}} \left( \frac{2\sigma^2}{2} \right) \sqrt{2\pi\sigma^2} = \sigma^2$$

(d)

(d.i) - 5 points

Instead of having just one Gaussian, we now have two Gaussians due to the fact that one fifth of the time the measurement apparatus adds 1 cm to the measurement. The probability of a length measurement  $d_i$  given the mean value is:

$$\mathbb{P}(d_i|L) = \frac{4}{5}\mathcal{N}(d_i|L, \sigma) + \frac{1}{5}\mathcal{N}(d_i|L+1, \sigma). \quad (7)$$

Let's rewrite this in terms of the exponential form of the Gaussian:

$$\mathbb{P}(d_i|L) = \frac{4}{5} \frac{c_1}{\sigma\sqrt{2\pi}} e^{-\frac{(d_i-L)^2}{2\sigma^2}} + \frac{1}{5} \frac{c_2}{\sigma\sqrt{2\pi}} e^{-\frac{(d_i-(L+1))^2}{2\sigma^2}}$$

We need to normalize this result so that the integral of this over all data (from  $-\infty$  to  $\infty$ ) is 1.

$$1 = \frac{4}{5} \frac{c_1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(d_i-L)^2}{2\sigma^2}} dL + \frac{1}{5} \frac{c_2}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(d_i-(L+1))^2}{2\sigma^2}} dL$$

These integrals are the same as what we already solved in part c. The addition of the  $d_i$  doesn't change the outcome of the integral, so we already know that both integrate to  $\sqrt{2\pi}\sigma$ . Thus we are left with:

$$1 = \frac{4}{5}c_1 + \frac{1}{5}c_2$$

Thus, both  $c_1$  and  $c_2$  are 1 and our distribution is normalized.

$$\mathbb{P}(d_i|L) = \frac{4}{5} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(d_i-L)^2}{2\sigma^2}} + \frac{1}{5} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(d_i-(L+1))^2}{2\sigma^2}}$$

See attached plot.

(d.ii) - 5 points

The probability of  $L$  given the data  $d_i$  is

$$\mathbb{P}(L|d_i) = \frac{4}{5}\mathcal{N}(L|d_i, \sigma) + \frac{1}{5}\mathcal{N}(L|d_i-1, \sigma). \quad (8)$$

We know from Bayes' Theorem that  $\mathbb{P}(L|d_i) \propto \mathbb{P}(d_i|L)P(L)$ . In this problem we have no prior guess, so the prior is a constant. This means that  $P(L)$  is a constant and can be eliminated from the proportion. Thus, in this case  $P(L|d_i) = P(d_i|L)$ . However, in order to plot the probability we will need to switch the variables  $d_i$  and  $L$ . In doing this substitution we must be careful to replace the  $(L+1)$  with  $(d_i-1)$  because the mean is one fifth of the time 1 cm less than the data point measured.

$$\mathbb{P}(d_i|L) = \frac{4}{5} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(L-d_i)^2}{2\sigma^2}} + \frac{1}{5} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(L-(d_i-1))^2}{2\sigma^2}}$$

This is going to give us a similar plot to part c, but the plot will appear mirrored around the data point (see attached plot).

**(d.iii) - 5 points**

Assuming the measurements are independent of each other. First calculate  $\mathbb{P}(d_1, d_2|L) = \mathbb{P}(d_1|L) \mathbb{P}(d_2|L)$ . Then interpret that as  $\mathbb{P}(L|d_1, d_2)$ , i.e., the posterior probability assuming a prior that equals 1 everywhere. Take the derivative with respect to  $L$  and set it zero to find the maximum probability  $L_{\max}$ . However, the product of the probabilities will have 3 local maxima. Plotting the expression is the easiest way to see where the global maximum is. If we plot the product of the two probabilities we will find that the graph has 3 peaks ( $L$  8.3,  $L$  9.25,  $L$  10.2). The maximum should be the peak at about  $L$  9.25.

$$\mathbb{P}(d_1, d_2|L) = \left( \frac{4}{5} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(d_1-L)^2}{2\sigma^2}} + \frac{1}{5} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(d_1-(L+1))^2}{2\sigma^2}} \right) \left( \frac{4}{5} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(d_2-L)^2}{2\sigma^2}} + \frac{1}{5} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(d_2-(L+1))^2}{2\sigma^2}} \right)$$

**Problem 3 - (graded by Dunzhu)****(a) - 5 points**

$$P(m) \propto e^{-F(m)}$$

$$F(m) = \sum_k \frac{1}{2} \frac{(d_k - g_k(m))^2}{\sigma_k^2}$$

**(b) - 5 points**

$W$  is a diagonal matrix, with  $W_{kk} = 1/\sigma_k$ .

**(c) - 5 points**

since

$$F(m) = \frac{1}{2} \sum_k (d'_k - g'_k(m))^2$$

This problem in terms of new data  $d'$  and new formula  $g'(m)$  can be solved using the formula in homework2.

But now

$$G'_{ik} = \frac{\partial g'_i}{\partial m_k} = \frac{1}{\sigma_i} \frac{\partial g_i}{\partial m_k}$$

$$\gamma' = -(G')^T (d' - g')$$

So

$$G' = WG$$

$$\gamma' = -G^T W^T W (d - g)$$

and the updating formula using inexact Hessian will be

$$m := m - ((G')^T G')^{-1} \gamma'$$

$$= m + (G^T W^T W G)^{-1} (G^T W^T W) (d - g)$$

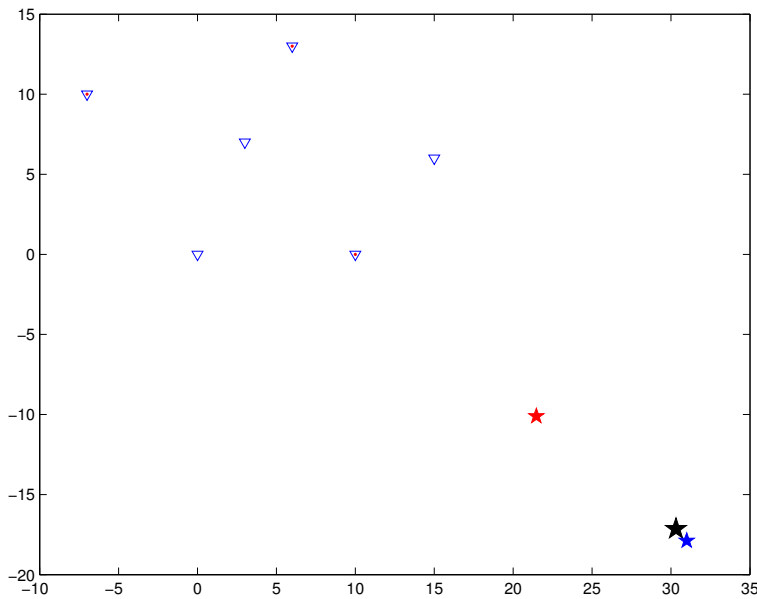


Figure 1: 3(d): Location of earthquake in three cases. Black star, case 1; blue star, case 2; red star, case 3.

#### (d) - 5 points

I considered three cases, where the first two case, earthquake location on surface is

$$\begin{aligned}\sigma &= [0.01, 0.01, 0.01, 0.01, 0.01, 0.01] \\ \sigma &= [0.007, 0.01, 0.01, 0.007, 0.1, 0.01] \\ \sigma &= [0.007, 9E4, 0.01, 9E4, 9E4, 0.01]\end{aligned}$$

The resulting model and misfit  $d_k - g_k(m)$  is

three models (ts, xs, ys, zs, v), the first two case, location on surface is similar, but depth is quite different. The third case, having a quite unrealistic sigma, which gives quite different location.

315.1474	315.8502	315.3964
30.3125	31.0006	21.4719
-17.1482	-17.8870	-10.1069
15.9867	9.8162	16.7610
5.2699	5.6505	4.1378

and misfit. Note misfit at k station is large when sigma\_k is large.

0.0010	0.0002	0
0.0026	0.0011	-0.1518
-0.0029	-0.0012	0.0000
0.0051	0.0011	0.1504
-0.0010	-0.0440	0.3286
-0.0048	-0.0021	0

```
1 function hw3()
2 xi=[0 10 15 6 -7 3]';
```

```

3 yi=[0 0 6 13 10 7]';
4 zi=[0 0 0 0 0 0]';
5 ti=[322.418 321.031 321.228 323.093 324.415 322.706]';
6
7 sigma0=zeros(6,1)+0.01;
8 M0=do_one_ti(xi,yi,zi,ti,sigma0);
9 ti_0=predict(xi,yi,zi,M0);
10
11 sigma1=[0.007, 0.01, 0.01, 0.007, 0.1, 0.01]';
12 M1=do_one_ti(xi,yi,zi,ti,sigma1);
13 ti_1=predict(xi,yi,zi,M1);
14
15 sigma2=[0.007, 90000.01, 0.01, 90000.007, 90000.1, 0.01]';
16 M2=do_one_ti(xi,yi,zi,ti,sigma2);
17 ti_2=predict(xi,yi,zi,M2);
18
19 plot(xi,yi,'bv'); hold on;
20 plot(xi([2,4,5]),yi([2,4,5]),'r. ');
21 plot(M0(2),M0(3),'kp','markersize',14,'markerfacecolor','k');
22 plot(M1(2),M1(3),'bp','markersize',10,'markerfacecolor','b');
23 plot(M2(2),M2(3),'rp','markersize',10,'markerfacecolor','r');
24 hold off;
25 print -depsc2 fig1.eps
26 [M0, M1, M2]
27 [ti_0-ti, ti_1-ti, ti_2-ti]
28
29
30 function M=do_one_ti(xi,yi,zi,ti,sigma)
31
32 M=[300 20 -10 10 2]'; % initial guess
33 for step = 1:1000
34     grad=zeros(5,1);
35     hess=zeros(5,5);
36     for i=1:length(xi)
37         [tmp_grad, tmp_hess]=get_grad_hessian(xi(i),yi(i),zi(i),ti(i),M);
38         grad = grad + tmp_grad/sigma(i)^2;
39         hess = hess + tmp_hess/sigma(i)^2;
40     end
41     err(step) = norm(predict(xi,yi,zi,M)-ti);
42     if(step > 1 && err(step) > err(step-1))
43         break;
44     end
45     M = M- pinv(hess)*grad;
46 end
47 plot(err);
48
49 function [grad,hess] = get_grad_hessian(xi,yi,zi,ti,M)
50 ts=M(1); xs=M(2); ys=M(3); zs=M(4); v=M(5);
51 R=sqrt((xs-xi)^2 + (ys-yi)^2 + (zs-zi)^2);
52 nx=(xs-xi)/R;

```

```
53 ny=(ys-yi)/R;
54 nz=(zs-zi)/R;
55 e=(ts+R/v-ti);
56
57 grad = e*[1 nx/v ny/v nz/v -R/v^2]';
58 sen = [1 nx/v ny/v nz/v -R/v^2]';
59 hess = sen*sen'; % approximate hess
60
61 function r=predict(xi,yi,zi,M)
62
63 ts=M(1); xs=M(2); ys=M(3); zs=M(4); v=M(5);
64 r=xi*0;
65 for i=1:length(xi)
66     R=sqrt((xs-xi(i))^2 + (ys-yi(i))^2 + (zs-zi(i))^2);
67     r(i)=(ts+R/v);
68 end
```