Solution to analysis in Home Assignment 1

Kalle Funck (kallef)

Analysis

In this report I will present my independent analysis of the questions related to home assignment 1. I swear that the analysis written here are my own. I have not discussed it with anyone.

1 Properties of random variables

a) Showing that the expected value of a normally distributed variable is μ is done below.

$$\mathcal{N}(x, \sigma, \mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \Rightarrow$$

$$E(x) = \int_{-\infty}^{\infty} x \cdot \mathcal{N}(X, \sigma, \mu)\, dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x \cdot \exp\left(\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx$$

$$t = \frac{x-\mu}{2\sigma},\ dt = \frac{dx}{2\sigma} \Rightarrow$$

$$E(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (2\sigma t + \mu) \cdot \exp\left(\frac{-1}{2}(2t)^2\right) 2\sigma\, dt$$

$$= \frac{4\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} te^{-2t^2}\, dt + \frac{2\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-2t^2}\, dt$$

Integration by parts knowing the integral of e^{-2t^2} being $\sqrt{\pi}/2$ gives the following:

$$= \frac{4\sigma}{\sqrt{2\pi}} \left(t\sqrt{\frac{\pi}{2}} - \int_{-\infty}^{\infty} \sqrt{\frac{\pi}{2}}\, dt\right) + \frac{2\mu}{\sqrt{2\pi}}\sqrt{\frac{\pi}{2}}$$

$$= \frac{4\sigma}{\sqrt{2\pi}} \left(t\sqrt{\frac{\pi}{2}} - t\sqrt{\frac{\pi}{2}}\, dt\right) + \mu$$

$$E(x) = \frac{4\sigma}{\sqrt{2\pi}} \cdot 0 + \mu = \mu$$

Then the variance is shown below:

$$E((x-\mu)^2) = E(x^2 - 2x\mu + \mu^2)$$

$$= E(x^2) - 2\mu E(x) + \mu^2 = E(x^2) - \mu^2$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 \cdot \exp\left(\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx - \mu^2$$

$$t = \frac{(x-\mu)^2}{2\sigma},\ dt = \frac{2x\,dx}{2\sigma} \Rightarrow$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (2\sigma t + \mu)^2 \cdot \exp\left(\frac{-1}{2}(2t)^2\right) 2\sigma\, dt - \mu^2$$

$$= \frac{\sqrt{2}}{\sqrt{\pi}} \int_{-\infty}^{\infty} (2\sigma t + \mu)^2 e^{-2t^2}\, dt - \mu^2$$

$$= \sqrt{\frac{2}{\pi}} \left(\int (2\sigma t)^2 e^{-2t^2} dt + \int 4\sigma t \mu e^{-2t^2} dt + \int \mu^2 e^{-2t^2} dt \right) - \mu^2$$

$$\int \mu^2 e^{-2t^2} dt = \mu^2 \sqrt{\frac{\pi}{2}}$$

$$Var(X) = \sqrt{\frac{2}{\pi}} \left(\int (2\sigma t)^2 e^{-2t^2} dt + \int 4\sigma t \mu e^{-2t^2} dt \right) + \mu^2 - \mu^2$$

$$= \sqrt{\frac{2}{\pi}} \left(\int (2\sigma t)^2 e^{-2t^2} dt + \int 4\sigma t \mu e^{-2t^2} dt \right)$$

$$= \sqrt{\frac{2}{\pi}} \left(\int (2\sigma t)^2 e^{-2t^2} dt + 4\sigma\mu \left[\frac{-1}{4} e^{-2t^2} \right]_{-\infty}^{\infty} \right)$$

Since t^2 always is positive the function e^-kt^2 will converge to zero in both positive and negative infinity.

$$= \sqrt{\frac{2}{\pi}} \int (2\sigma t)^2 e^{-2t^2} dt = 4\sigma^2 \sqrt{\frac{2}{\pi}} \int t \cdot t e^{-2t^2} dt$$

Integration by parts gives us the following:

$$= 4\sigma^2 \sqrt{\frac{2}{\pi}} \left(\left[t \cdot \frac{-e^{-2t^2}}{4} \right]_{-\infty}^{\infty} - \int \frac{-e^{-2t^2}}{4} dt \right)$$

Convergence to zero for the first term.

$$Var(x) = \sigma^2 \sqrt{\frac{2}{\pi}} \int e^{-2t^2} dt = \sigma^2 \sqrt{\frac{2}{\pi}} \sqrt{\frac{\pi}{2}} \Rightarrow$$

$$Var(x) = \sigma^2$$

b) Firstlt the mean is shown below:

$$E(q) = \int_{-\infty}^{\infty} q\, p(q)\, dq$$

$$E(z) = \int_{-\infty}^{\infty} z\, p(q)\, dq = \int_{-\infty}^{\infty} Aq\, p(q)$$

$$E(z) = AE(q)$$

To then showcase the covariance:

$$Cov(z) = E\left[(z - E[z])(z - E[z])^T\right]$$

$$= E\left[(Aq - AE[q])(Aq - AE[q])^T\right]$$

$$= AE[(q - E[q])(A(q - E[q]))^T]$$

$$= AE[(q - E[q])([q - E[q]])^T A^T] = ACov(q)A^T$$

c) This task was done in MATLAB and yielded the following figure:

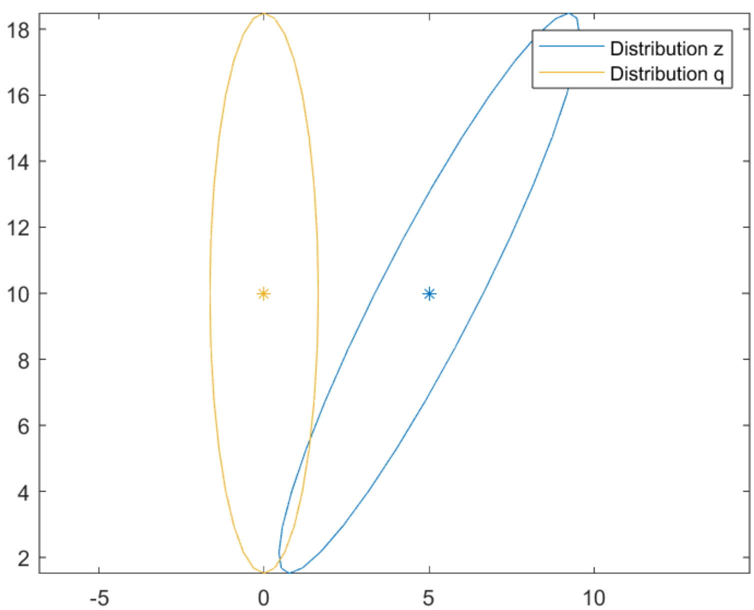

Figure 1.1: 2d distrubutions for variables z and q

	μ	σ
z	$\begin{bmatrix} 5 \\ 10 \end{bmatrix}$	$\begin{bmatrix} 2.3000 & 4.0000 \\ 4.0000 & 8.0000 \end{bmatrix}$

Table 1: Distribution values of z

Because of $p(q)$ does not have any correlation at all, the non-diagonals are set to zero you can expect the form that is shown where it is straight. However, as you add in the A matrix you suddenly get a positive correlation and it should therefore start pointing positively diagonally as it does. Which the covariance matrix also shows. The mean has been altered to move to the right which goes along with this altered covariance since now the x-value gets affected by y positively and therefore pushes the mean to the right.

2 Transformation of random variables

a) To determine $p(z)$ this analytically the `affineGaussianTransform` function was used and with the A "matrix" as 3. To then approximate the Gaussian distribution `approxGaussianTransform` was used with different steps to give a feeling of what happens when the number of samples are altered. Which resulted in the following:

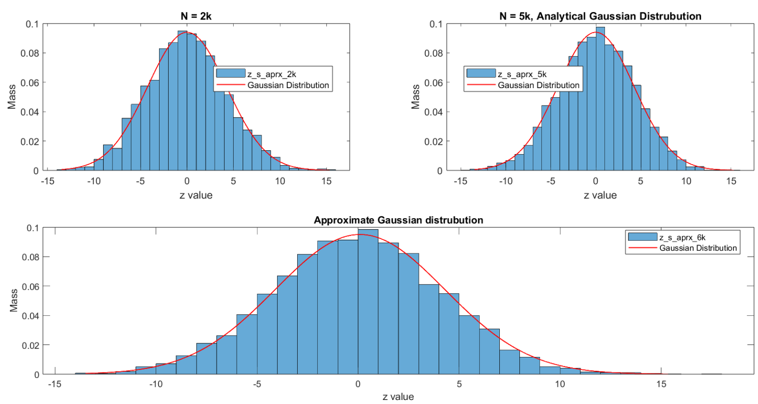

Figure 2.1: Comparison between analytical and approximal generation of the stochastic variable z for different values of samples.

σ_z^2	μ_z
18	0

Table 2: Distribution values $p(z)$

	$N = 2000$	$N = 5000$
σ_z^2	18.8201	17.6731
μ_z	0.0486	0.0073

Table 3: Distribution values for the approximated curves

As can be seen in the figure they match pretty well even though they have some differences. By doing this random distribution you can draw the conclusion that this will be enough for many purposes and when the distribution is unknown this is a good way of estimating it. It must be concluded however that the analytical is more precise and also faster to compute. To achieve a higher and more precise distribution the variable N can be increased since it will allow the approximated curve to be closer and closer to the reality. Different experiments with different N provided that insight since lower N tend to be more random and if you increase the number of samples the closer to the real Gaussian distribution you come. Therefore if you want precision you will want to increase N but it will decrease computation speed which can potentially be a problem in real-time systems. As for the two pdfs, they are very similar and could for many purposes both be used but if you require very detailed precision the analytical gives more of that.

b) In this case the function is nonlinear which means that we can not assume that the following density will be a Gaussian distribution itself and the function we implemented earlier for affine transformation will not work. Therefore only the approximated Gaussian will be shown here. N has the same effect as earlier which was described and is therefore not shown. The more samples the more alike it will be a normal distribution. The provided results are shown here:

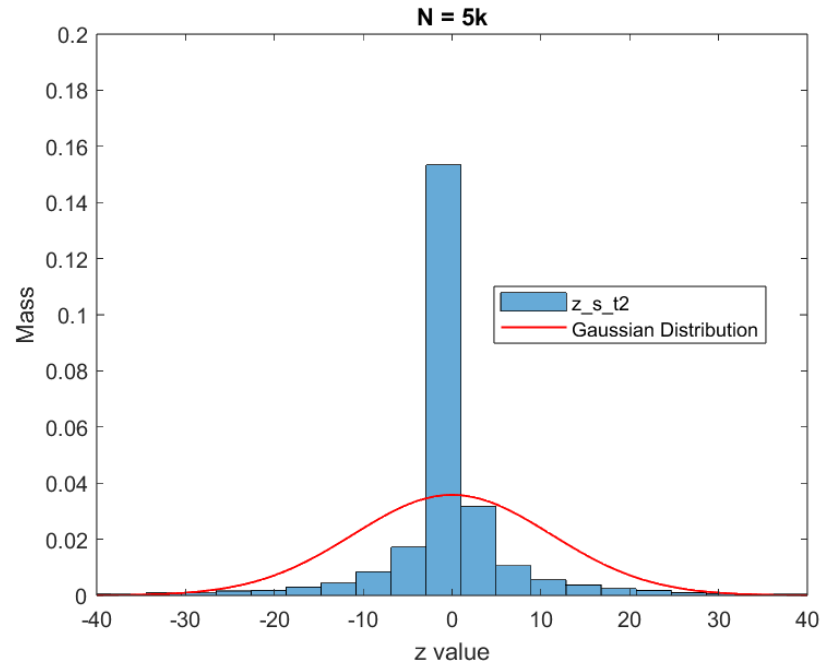

Figure 2.2: Gaussian approximation of the distribution of $z = x^3$

σ_z^2	μ_z
124.2526	0.0143

Table 4: Distribution values $p(z)$

As can be seen, the samples generated do not match as well with the Gaussian distribution it predicts and this is because the standard deviation gets affected by the increased occurrences of really large numbers. Which definitely occur when you take the cubic function of a dataset. Three standard deviations away from the average would take you to ± 6 in the original distribution but if you cube that it will be ± 216. Which goes to show that the frequency of extreme numbers will spread. However, if the answer is between -1 and 1 it will become closer to zero which results in that the majority of the previous distribution will stay there. The histogram and the pdf does not match too well for this function but if the function was totally unknown it could help with predicting were it will land even though the estimated curve might not be too accurate.

c) If we were to compare these results we can firstly conclude that estimating a Gaussian distribution works better when the distribution is in fact normally distributed, as it is in question a). However, sometimes it might be suitable to adjust the distribution and act as if it has a Gaussian distribution as in question b). If you were to add it to a Kalman filter you would want to assume Gaussian noise as an example and it is therefore suitable approximate a Gaussian distribution in some cases even though it might not be distributed that way in reality.

3 Understanding the conditional density

a) Because that we have no information on x I would say that it is impossible to conclude anything about where y will end. Unless we get the distrubution of x all values of x are equally probable and hence the value of $h(x)$ will have the same issue. And the final part of this chain is that y also get affected in the same way since we do not know the probability of $h(x)$. Therefore we must know more to be able to tell the probability of y.

b) It is now possible to describe that distribution since we have the information we lacked in the previous question. Since $h(x)$ is deterministic we know that sending a known x into the function will result in a known constant. Since r have a Gaussian distribution, adding a constant to it will simply alter the mean. Meaning y will have a distribution of $\mathcal{N}(h(x), \sigma_r^2)$. Where $h(x)$ is a known constant.

c) We have the same issue here as in the previous questions, if we do not have any information on x it is not possible to generate any statistical probability on where y will end up. As far as we know x could range anywhere between negative infinity to infinity, risking $h(x)$ and y to do the same.

   If we however know x the mean simply get altered like earlier where $y \sim \mathcal{N}(Hx, \sigma_r^2)$

d) In this case we have information about x and we begin the discussion with looking over the non-linear example of $h(x)$. In this case we can not assume that the distribution that comes out of that function has a Gaussian distribution as we saw in the previous section, where a function was $z(x) = x^3$ and the result changed the distribution. It might be possible to describe $p(y)$ but we can not assume it will be Gaussian. I would need to have the function to be able to tell what the distribution would look like but it would probably be some kind of different distribution that is not used too frequently. If x is given, as before the following distribution will occur: $y \sim \mathcal{N}(h(x), \sigma_r^2)$.

   Let us talk about if $h(x) = Hx$. In this case we can assume $h(x)$ will be normally distributed as well and therefore y as well since adding two

Gaussians will result in a Gaussian. To determine what the values will be it is not as clear in this case so the following math is done.

$$\mu_y = E[y] = E[Hx + r] = HE[x] + E[r] = H\mu_x$$

$$\sigma_y^2 = E[(y - E[y])^2] = E[(Hx + r - H\mu_x)^2]$$

$$= E[((Hx - H\mu_x) + (r - 0))^2]$$

$$= E[(Hx - H\mu_x)^2] + E[2r(Hx - H\mu_x)] + E[r^2 - 0]$$

Independent x and r gives us ability to remove the cross-term.

$$= E[(Hx - H\mu_x)^2] + E[r^2 - 0] = H^2\sigma_x^2 + \sigma_r^2$$

$$y \sim \mathcal{N}(H\mu_x, H^2\sigma_x^2 + \sigma_r^2)$$

e) Then it was left to try and analyze this through numerical experiments, firstly what was done was setting the variables my self which can be viewed in the box below.

σ_r	h(x)	H	σ_x	μ_x
2	e^x	5	3	1

Table 5: Caption

Then it was tested to see what happens when $h(x) = e^x$ and also $h(x) = Hx$ for a uniformly distributed variable x and for fixed x, which was set as $x = 1$. The probability $p(y|x)$ changes if you change x so in this case it is only actual for that given x but it is also possible to change it to get a new probability. The following results got provided for the four different tests:

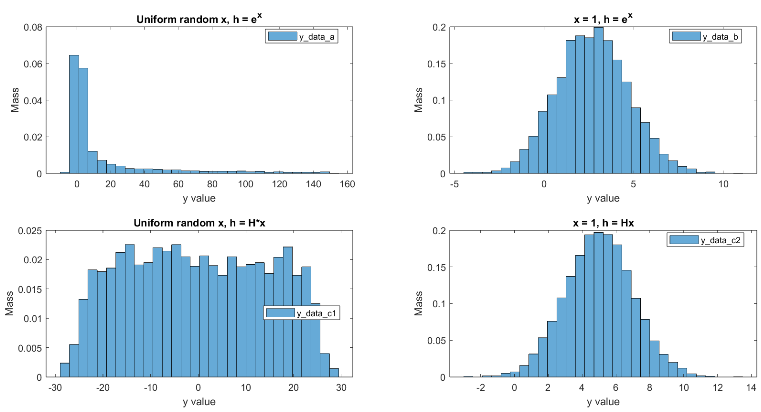

Figure 3.1: Random distributions for different h(x) and for when x is determined and when it is not

As can be seen when when x is randomly distributed it does not look like a Gaussian at all which is reasonable as the argumentation in the previous tasks tell. When the x is fixed it however looks like the curves described earlier. In the top left graph you can see that the mean of the curve is around 2.75 which is close to e which it should. As for the bottom one it is located at 5 which is also what it should be with the reasoning above.

As for the Gaussian distributed x the result became the following:

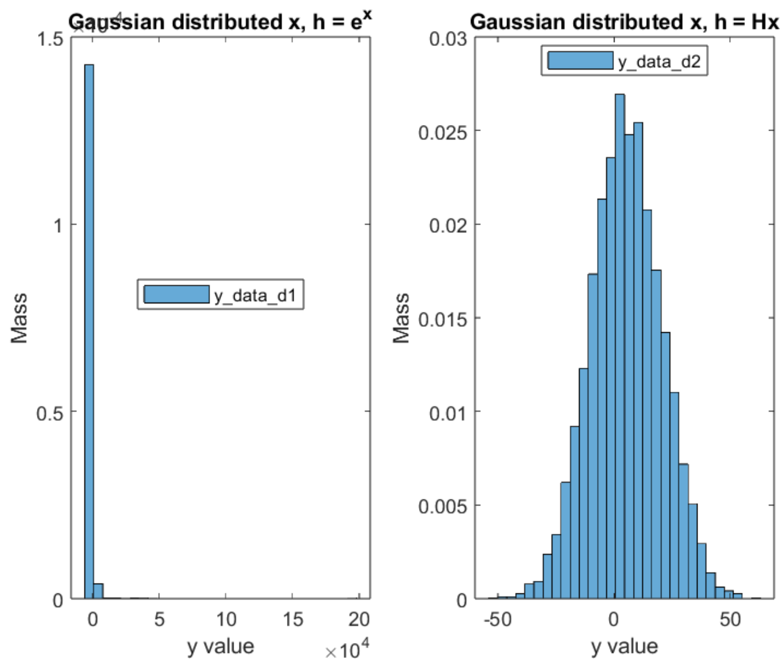

Figure 3.2: Random distributions for different h(x) for when input x is normally distributed

Showcased in the image on the left the distribution does not appear to be Gaussian and that is reasonable since $h(x)$ does transform it away from being a normal distribution. However the subfigure on the right seems to be a Gaussian distribution. The mean was tested in MATLAB and was 5.3 which is at least close to 5. Standard deviation was 15.18 which should be equal to $\sqrt{H^2\sigma_x^2 + \sigma_r^2} = 15.13$. Which highlights that the samples behaves as it should and the reasoning in task d) is sound. This test can not prove the earlier stated facts but it can at least show us that there is not any clear errors with the earlier reasoning since these experiments concludes the same as the analyze did.

4 MMSE and MAP estimators

Implementing this in MATLAB provided the following results:

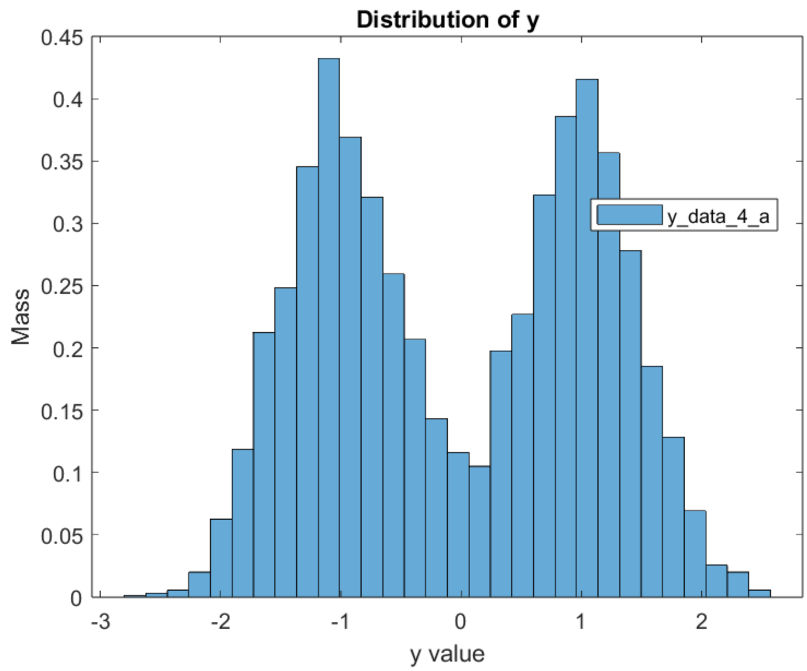

Figure 4.1: Distribution of $y = \theta + w$, samples = 5000

a) As can be seen there are two medians centered around -1 and 1 which is rational since the measurement error w is centered around 0 and does not push the medians from θ away. The noise is quite high it seems like since the two medians intersect around 0. If you were to classify which of these measurements are -1 and 1 you would not do it perfectly since a lot are centered around 0. Hence the sensor disturbances are a bit too high.

b) Intuitively 0.7 should be classified as a 1 because it is far closer but let us look at this mathematically. 0.7 is $0.3/0.5 = 0.6$ standard deviations away from the pillar centered at around 1. It is $1.7/0.5 = 3.4$ standard deviations away from the center of the curve located at -1. The chance of a variable being 3.4σ away or more from the mean is equal to 0.07%. (Calculated using digital tools). However, the chance of a variable being 0.6σ away or more from the mean is equal to 54.85%. Therefore it is extremely likely that the real θ was 1 in this case.

c) Firstly to determine $p(y|\theta)$ we can quite simply see that if we know θ it becomes a constant and therefore alters the mean of the final distrubution. The variance becomes dependent on the variance of w. Therefore, $p(y|\theta) \sim \mathcal{N}(\theta, \sigma^2)$.

To then conclude $p(y)$ the following calculations was done.

$$p(y) = \int p(y|\theta)\, p(\theta)\, d\theta$$

Since we only have two possibilites of θ this can be split into

$$p(y) = \int p(y|\theta_1)\, p(\theta_1) d\theta_1 + \int p(y|\theta_2)\, p(\theta_2)\ d\theta_2$$

Since $p(\theta_{1,2})$ are both impulse function meaning that integrating over them only gives the possibility. $p(\theta_1) = 0.5\delta(\theta - 1)$ as an example and integrating over it will only provide 0.5 over it's whole domain. Since $\theta_{1,2}$ both are known they can be addid into the normal distrubution meaning that $p(y|\theta_{1,2}$ will not have any θ left to integrate over. Mathematically this can be expressed as:

$$\begin{aligned} p(y) &= \int 0.5\mathcal{N}(1, \sigma)\, \delta(\theta_1 - 1)\, d\theta_1 + \int 0.5\mathcal{N}(-1, \sigma)\, \delta(\theta_2 + 1)\, d\theta_2 \\ &= 0.5\mathcal{N}(1, \sigma) + 0.5\mathcal{N}(-1, \sigma) \\ &= 0.5\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y-1)^2}{2\sigma^2}\right) + 0.5\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y+1)^2}{2\sigma^2}\right) \end{aligned}$$

Which is what is what was supposed to be shown.

d) The bayesian rule states:

$$p(\theta|y) = \frac{p(y|\theta)\, p(\theta)}{p(y)}$$

$$p(\theta = 1|y) = \frac{p(y|\theta = 1)\, p(\theta = 1)}{p(y)}$$

Putting in the known values from before provides the following:

$$p(\theta = 1|y) = \frac{\mathcal{N}(1, \sigma) \cdot 0.5}{0.5(\mathcal{N}(1, \sigma) + \mathcal{N}(-1, \sigma))} = \frac{\mathcal{N}(1, \sigma)}{\mathcal{N}(1, \sigma) + \mathcal{N}(-1, \sigma)}$$

$$p(\theta = 1|y) = \frac{e^{\frac{-1}{2}\left(\frac{y-1}{\sigma}\right)^2}}{e^{\frac{-1}{2}\left(\frac{y-1}{\sigma}\right)^2} + e^{\frac{-1}{2}\left(\frac{y+1}{\sigma}\right)^2}}$$

Doing the same for $\theta = -1$ gives us the this.

$$p(\theta = -1|y) = \frac{\mathcal{N}(-1, \sigma)}{\mathcal{N}(1, \sigma) + \mathcal{N}(-1, \sigma)}$$

Which can then be expanded like shown above just switch the nominator from $y-1$ to $y+1$.

e) Using the following formula presented in the lab PM for our two θ this results in the following sum.

$$\hat{\theta} = 1\,\mathrm{Pr}(\theta = 1|y) + (-1)\,\mathrm{Pr}(\theta = -1|y)$$

$$\hat{\theta} = \frac{\mathcal{N}(1,\sigma)}{\mathcal{N}(1,\sigma) + \mathcal{N}(-1,\sigma)} - \frac{\mathcal{N}(-1,\sigma)}{\mathcal{N}(1,\sigma) + \mathcal{N}(-1,\sigma)}$$

$$= \frac{e^{\frac{-1}{2}\left(\frac{y-1}{\sigma}\right)^2}}{e^{\frac{-1}{2}\left(\frac{y-1}{\sigma}\right)^2} + e^{\frac{-1}{2}\left(\frac{y+1}{\sigma}\right)^2}} - \frac{e^{\frac{-1}{2}\left(\frac{y+1}{\sigma}\right)^2}}{e^{\frac{-1}{2}\left(\frac{y-1}{\sigma}\right)^2} + e^{\frac{-1}{2}\left(\frac{y+1}{\sigma}\right)^2}}$$

f) Since you try to optimize the function $p(\theta|y)$ you will get either one or negative one depending on the reading from y. This is because that the function can only be one of those measurements and depending which spike it is closer to, -1 or 1 it will return one of them. It is the value of θ that gives the highest probability of θ given y.

g) The difference between these two examples can be viewed by that MMSE is a contunious function which will not return the predicted value of θ according to the rules of this distrubution. It will probably work better on a variable that can not be one of two options. Trying to calculate this in MATLAB, addind $y = 0.7$ as the variable and then solving the equation returns $\hat{\theta} = 0.6567$. While that might be probable if the distrubution was not stacked around -1 and 1 and instead had a continuous distrubution, in this case it does not provide what we are searching for. What you can do is round it to the closest choice which is 1 in this case and then you can conclude that it works but the MAP works better for this situation.

Testing the MAP estimator yields $p(\theta = 1|y = 0.7) = 0.7595$ while $p(\theta = -1|y = 0.7) = 0.1028$. In this case the value of θ that provides the greatest value of $p(\theta|y)$ is $\theta = 1$.