ECE531 Homework Assignment Number 6 Solution

Due by 8:50pm on Wednesday 23-Mar-2011

Make sure your reasoning and work are clear to receive full credit for each problem.

1. 6 points. Suppose you have a scalar random parameter $\Theta \in \mathbb{R}$ with discrete prior distribution $\pi(\theta) = \frac{1}{2}(\delta(\theta-1) + \delta(\theta+1))$. You receive a single observation

$$Y = \Theta + W$$

where $W \sim \mathcal{N}(0, \sigma^2)$ and σ^2 is known.

(a) 2 points. Find the Bayesian MMSE estimator for the parameter Θ.

Solution: All of the Bayesian estimators require us to compute the posterior density

$$\pi_y(\theta) = \frac{p_\theta(y)\pi(\theta)}{p(y)}$$

First we calculate the unconditional density $p(y)$:

$$\begin{aligned} p(y) &= \int_\Lambda p_\theta(y)\pi(\theta)d\theta \\ &= \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}(y-1)^2\} + \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}(y+1)^2\} \end{aligned}$$

Then we can compute the posterior density:

$$\begin{aligned} \pi_y(\theta) &= \begin{cases} \frac{\frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}(y-1)^2\}}{\frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}(y-1)^2\}+\frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}(y+1)^2\}} & \text{if } \theta = 1 \\ \frac{\frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}(y+1)^2\}}{\frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}(y-1)^2\}+\frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}(y+1)^2\}} & \text{if } \theta = -1 \end{cases} \\ &= \begin{cases} \frac{\exp\{\frac{y}{\sigma^2}\}}{\exp\{\frac{y}{\sigma^2}\}+\exp\{-\frac{y}{\sigma^2}\}} & \text{if } \theta = 1 \\ \frac{\exp\{-\frac{y}{\sigma^2}\}}{\exp\{\frac{y}{\sigma^2}\}+\exp\{-\frac{y}{\sigma^2}\}} & \text{if } \theta = -1 \end{cases} \end{aligned}$$

The MMSE estimate is then the conditional mean,

$$\begin{aligned} \hat{\theta}_{MMSE}(y) &= \int_\Lambda \theta\pi_y(\theta)d\theta \\ &= \pi(\theta=1|y) - \pi(\theta=-1|y) \\ &= \frac{\exp\frac{y}{\sigma^2} - \exp-\frac{y}{\sigma^2}}{\exp\frac{y}{\sigma^2} + \exp-\frac{y}{\sigma^2}} \\ &= \frac{2\sinh\left(\frac{y}{\sigma^2}\right)}{2\cosh\left(\frac{y}{\sigma^2}\right)} = \tanh\left(\frac{y}{\sigma^2}\right) \end{aligned}$$

(b) 2 points. Find the Bayesian MAP estimator for the parameter Θ.

Solution: The MAP estimator is the conditional mode, which we can compute as

$$\begin{aligned}
\hat{\theta}_{MAP}(y) &= \arg\max_{\theta=\pm 1} \pi_y(\theta) \\
&= \begin{cases} 1 & \text{if } \frac{\exp\frac{y}{\sigma^2}}{\exp\frac{y}{\sigma^2}+\exp-\frac{y}{\sigma^2}} \geq \frac{\exp-\frac{y}{\sigma^2}}{\exp\frac{y}{\sigma^2}+\exp-\frac{y}{\sigma^2}} \\ -1 & \text{if } \frac{\exp\frac{y}{\sigma^2}}{\exp\frac{y}{\sigma^2}+\exp-\frac{y}{\sigma^2}} < \frac{\exp-\frac{y}{\sigma^2}}{\exp\frac{y}{\sigma^2}+\exp-\frac{y}{\sigma^2}} \end{cases} \\
&= \begin{cases} 1 & \text{if } y \geq 0 \\ -1 & \text{if } y < 0 \end{cases} \\
&= \operatorname{sgn}(y)
\end{aligned}$$

(c) 1 point. Under what conditions are the MMSE and MAP estimates approximately equal?

Solution: We can see that $\tanh(x) \to 1$ as $x \to \infty$ and similarly, $\tanh(x) \to -1$ as $x \to -\infty$. Thus, if σ^2 is very small ($\ll 1$), then the two estimators will be approximately equal.

(d) 1 points. Discuss how this problem relates to simple binary Bayesian hypothesis testing.

Solution: In this problem, θ can only take on two possible values. So, it could also be considered in the context of detection if we assign hypotheses $\mathcal{H}_1 : \theta = 1$ and $\mathcal{H}_0 : \theta = -1$. We know the Bayes decision rule in this case is to decide $\mathcal{H}_1$ when $y \geq 0$ and decide $\mathcal{H}_0$ when $y < 0$ (with ambiguity at $y = 0$).

The $\hat{\theta}_{MMSE} = \tanh(y/\sigma^2)$ MMSE estimator can be thought of as a "soft" decision rule in the sense that the estimate is always in $[-1, 1]$ and approaches one of the possible values for θ when the evidence in favor of $+1$ or -1 is very strong, i.e. when $y/\sigma^2 \to \infty$ or $y/\sigma^2 \to -\infty$. You can use the MMSE estimate to get a Bayes detector by deciding $\mathcal{H}_1 : \theta = 1$ when $\hat{\theta} \geq 0$ and deciding $\mathcal{H}_0 : \theta = -1$ otherwise. But the "soft" MMSE estimate also gives a result that reflects to some extent the uncertainty of the decision.

The MAP *estimator* gives the same result as a Bayesian *detector* in this problem (decide $\mathcal{H}_1 : \theta = 1$ when $y \geq 0$ and decide $\mathcal{H}_0 : \theta = -1$ otherwise).

2. 4 points. Kay I, problem 10.3.

Solution: This problem requires the computation of the posterior density $p_y(\theta)$ and then the conditional mean. I will use the notation y_n for the observations (Kay uses $x[n]$). Note that, since the observations are conditionally independent, the *joint* conditional density of the observations can be written as the product of the marginals

$$p_\theta(y) = \prod p_\theta(y_n) = \begin{cases} \exp\left(-\sum(y_n - \theta)\right) & \text{all } y_n > \theta \\ 0 & \text{otherwise.} \end{cases}$$

If we let $\check{y} = \min y_n$ and $\bar{y} = \frac{1}{N}\sum y_n$, we can write the joint conditional density as

$$p_\theta(y) = \begin{cases} \exp\left(-N(\bar{y} - \theta)\right) & \check{y} > \theta \\ 0 & \text{otherwise.} \end{cases}$$

The posterior density can then be computed as

$$\begin{aligned} \pi_y(\theta) &= \frac{p_\theta(y)\pi(\theta)}{p(y)} \\ &= \begin{cases} \frac{\exp(-N(\bar{y}-\theta))\exp(-\theta)}{\int_0^{\check{y}} \exp(-N(\bar{y}-\theta))\exp(-\theta)\,d\theta} & 0 < \theta < \check{y} \\ 0 & \text{otherwise.} \end{cases} \\ &= \begin{cases} \frac{(N-1)\exp(\theta(N-1))}{\exp(\check{y}(N-1))-1} & 0 < \theta < \check{y} \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Note that the posterior density is only a function of $\check{y} = \min y_n$ and N. All of the $\bar{y}$ terms cancelled out. Also note this result is only valid for $N \geq 2$. If $N = 1$, we can perform this same calculation with the marginal conditional density.

To compute the MMSE estimator, we need to compute the conditional mean. Note that only the numerator of $\pi_y(\theta)$ is a function of θ, so we can write

$$\begin{aligned} \hat{\theta}_{MMSE} &= \mathrm{E}[\theta \,|\, Y = y] \\ &= \frac{\int_0^{\check{y}} \theta(N-1)\exp(\theta(N-1))\,d\theta}{\exp(\check{y}(N-1)) - 1} \end{aligned}$$

The integral is of the form $\int xe^x\,dx$, which you solve via integration by parts (or a lookup table or a symbolic math software package). After performing the integral and doing some algebra to simplify the result, we get

$$\hat{\theta}_{MMSE} = \frac{\check{y}}{1 - \exp\left\{-\check{y}(N-1)\right\}} - \frac{1}{N-1}.$$

As a sanity check, let's see what happens when $N \to \infty$. In this case $\hat{\theta}_{MMSE} \to \check{y}$, which should make intuitive sense. As the number of observations becomes large, the prior is "washed away" and only the observations matter. The parameter θ represents the minimum value of the observations (the left edge of the conditional density), hence $\hat{\theta}_{MMSE} \to \check{y}$ makes intuitive sense.

When the number of observations is finite, the prior acts as a correction factor on the estimate in the sense that the data is optimally weighted with the prior to minimize the Bayes MMSE risk. The prior in this problem says that θ is more likely to be small than large. Figure 1 shows the MMSE estimate as a function of $\check{y}$ and N for values of $N = 2, 3, \ldots, 10$. As expected, the prior causes the MMSE estimate to always be less than $\check{y}$, and especially so for smaller values of N.

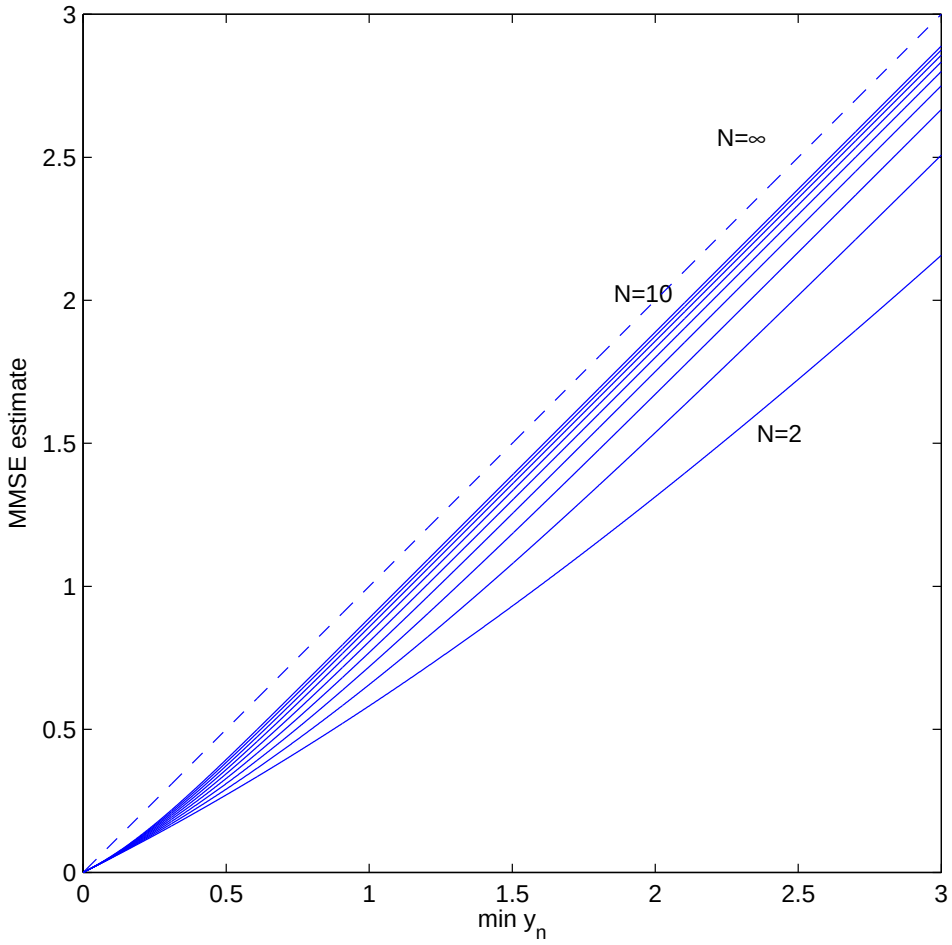

Figure 1: MMSE estimates in problem 2 as a function of $\check{y}$ and N.

3. 4 points. Kay I, problem 10.4.

Solution: As before, we need to compute the posterior density $p_y(\theta)$ and then compute the conditional mean. I will use the notation y_n for the observations (Kay uses $x[n]$). Since the observations are conditionally independent, the *joint* conditional density of the observations can be written as the product of the marginals

$$p_\theta(y) = \prod p_\theta(y_n) = \begin{cases} \frac{1}{\theta^N} & \text{all } y_n \text{ satisfy } 0 \le y_n \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\min y_n \ge 0$ for any parameter θ in this problem since $\Theta \sim \mathcal{U}[0, \beta]$. So, if we let $\hat{y} = \max y_n$, we can write the joint conditional density as

$$p_\theta(y) = \begin{cases} \frac{1}{\theta^N} & \hat{y} \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

Also note that

$$\pi(\theta) = \begin{cases} \frac{1}{\beta} & 0 \le \theta \le \beta \\ 0 & \text{otherwise.} \end{cases}$$

We have everything we need. The posterior density can then be computed as

$$\begin{aligned} \pi_y(\theta) &= \frac{p_\theta(y)\pi(\theta)}{p(y)} \\ &= \begin{cases} \frac{\frac{1}{\theta^N}\frac{1}{\beta}}{\int_{\hat{y}}^{\beta} \frac{1}{\theta^N}\frac{1}{\beta}\, d\theta} & \hat{y} \le \theta \le \beta \\ 0 & \text{otherwise.} \end{cases} \\ &= \begin{cases} \frac{\frac{1}{\theta^N}}{\frac{1}{N-1}\left(\hat{y}^{-(N-1)} - \beta^{-(N-1)}\right)} & \hat{y} \le \theta \le \beta \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Note this result is only valid for $N \ge 2$. If $N = 1$, we can perform this same calculation with the marginal conditional density. To compute the MMSE estimator, we need to compute the conditional mean. Note that only the numerator of $\pi_y(\theta)$ is a function of θ, so we can write

$$\begin{aligned} \hat{\theta}_{MMSE} &= \mathrm{E}[\theta \,|\, Y = y] \\ &= \frac{\int_{\hat{y}}^{\beta} \theta \frac{1}{\theta^N}\, d\theta}{\frac{1}{N-1}\left(\hat{y}^{-(N-1)} - \beta^{-(N-1)}\right)} \\ &= \frac{\hat{y}^{-(N-2)} - \beta^{-(N-2)}}{\hat{y}^{-(N-1)} - \beta^{-(N-1)}} \cdot \frac{N-1}{N-2} \end{aligned}$$

Note this result is only valid for $N \ge 3$.

When the number of observations is finite, the prior acts as a correction factor on the estimate in the sense that the data is optimally weighted with the prior to minimize the Bayes MMSE risk. Since $\hat{y}$ is always going to be less than the parameter of interest, we can expect the correction to cause the estimate to be larger than $\hat{y}$. Figure 2 shows the MMSE estimate as a function of $\hat{y}$ and N when $\beta = 1$ for values of $N = 3, 4, \ldots, 10$. As expected, the prior causes the MMSE estimate to always be more than $\hat{y}$, and especially so for smaller values of N.

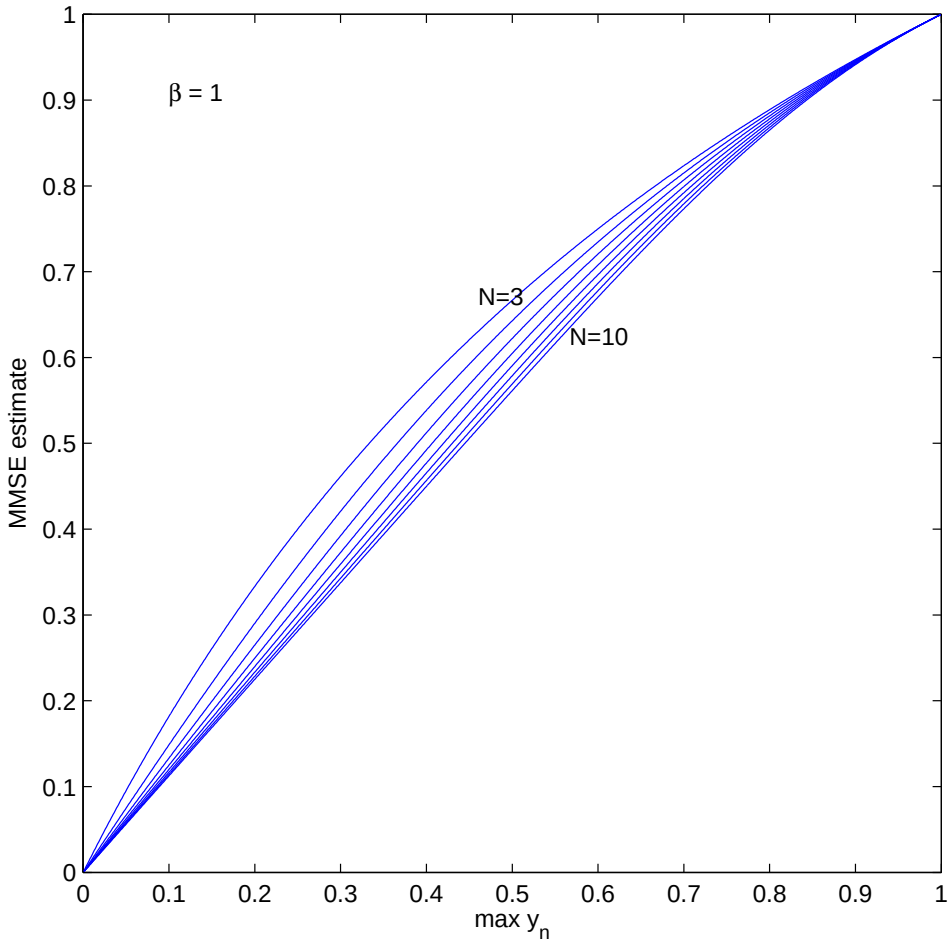

Figure 2: MMSE estimates in problem 3 as a function of $\hat{y}$ and N for the case when $\beta = 1$.

When β is large and $N \geq 3$, we can write

$$\begin{aligned}\hat{\theta}_{MMSE} &\approx \frac{\hat{y}^{-(N-2)}}{\hat{y}^{-(N-1)}} \cdot \frac{N-1}{N-2} \\ &= \hat{y}\frac{N-1}{N-2}.\end{aligned}$$

If N is also large, then $\hat{\theta}_{MMSE} \to \hat{y}$. This makes intuitive sense since the observations are upper bounded by θ and the maximum observation should give a good estimate for θ as $N \to \infty$.

4. 4 points. Kay I, problem 10.12.

Solution: Everything here is jointly Gaussian (or bivariate Gaussian since there are only two random variables). If we let $z = [x, y]^\top$, the joint density can be written as

$$p_Z(z) = p_{X,Y}(x, y) = \frac{1}{2\pi|C|^{1/2}} \exp\left\{\frac{-z^\top C^{-1} z}{2}\right\}.$$

Now conditioning on $x = x_0$ and doing a bit of algebra, we have

$$g(y) = \frac{1}{2\pi|C|^{1/2}} \exp\left\{\frac{-(x_0^2 - 2\rho x_0 y + y^2)}{2(1-\rho^2)}\right\} = \frac{1}{2\pi|C|^{1/2}} \exp\left\{\frac{-h(y)}{2(1-\rho^2)}\right\}.$$

So $g(y)$ is maximized when $h(y)$ is minimized. To minimize $h(y)$, we take a derivative and set the result to zero, i.e.

$$\frac{d}{dy} h(y) = 2y - 2\rho x_0 = 0$$

which has a unique solution at $y = \rho x_0$. You can take another derivative to confirm that this is indeed a minimum of $h(y)$. Hence, we have proven that $g(y)$ is maximized when $y = \rho x_0$.

Now, we can use the results in Theorem 10.1 of Kay I for bivariate Gaussian distributed random variables to write

$$\mathrm{E}[Y \mid X = x_0] = \mathrm{E}[Y] + \frac{\mathrm{cov}(X, Y)}{\mathrm{var}(X)}(x_0 - \mathrm{E}[X]).$$

From the problem description, we know $\mathrm{E}[Y] = \mathrm{E}[X] = 0$, $\mathrm{cov}(X, Y) = \rho$, and $\mathrm{var}(X) = 1$. Hence,

$$\mathrm{E}[Y \mid X = x_0] = \rho x_0.$$

Why are they the same? Conditioning on $X = x_0$ does not change the Gaussianity of Y. All the conditioning does is make the posterior distribution of Y have a different mean and variance, but the posterior distribution of Y remains Gaussian. We know that the symmetry of a Gaussian distribution causes the mean, median, and mode to all be the same thing.

When $\rho = 0$, X and Y are uncorrelated (actually independent since uncorrelated Gaussian random variables are independent). So observing $X = x_0$ tells us nothing about Y. As you would expect then,

$$\begin{aligned}
\mathrm{E}[Y \mid X = x_0] &= \mathrm{E}[Y] + \frac{\mathrm{cov}(X, Y)}{\mathrm{var}(X)}(x_0 - \mathrm{E}[X]) \\
&= \mathrm{E}[Y] + 0 \cdot (x_0 - \mathrm{E}[X]) \\
&= \mathrm{E}[Y] \\
&= 0.
\end{aligned}$$

5. 3 points. Kay I, problem 11.3. Also find the Bayesian MMAE estimator.

Solution: We've already been given the posterior pdf in this problem, so we just need to compute the conditional mean, median, and mode to get the MMSE, MMAE, and MAP estimators, respectively. I will use the notation y for my observation (Kay uses x). We can compute

$$\begin{aligned}\hat{\theta}_{MMSE}(y) &= \mathrm{E}[\Theta \,|\, Y = y] \\ &= \int_y^\infty \theta \exp(-(\theta - y))\, d\theta \\ &= y + 1\end{aligned}$$

This makes sense because the posterior distribution is an exponential distribution with parameter $\lambda = 1$ shifted to the right by y. The mean of an exponentially distributed random variable with parameter λ is $\frac{1}{\lambda}$. Hence, the mean of this shifted exponential distribution should be $y + \frac{1}{\lambda} = y + 1$.

The median of an exponentially distributed random variable with parameter λ is $\frac{\ln 2}{\lambda}$. What we have here is an exponential distribution with parameter $\lambda = 1$ shifted to the right by y. Hence,

$$\begin{aligned}\hat{\theta}_{MMAE}(y) &= \mathrm{median}[\Theta \,|\, Y = y] \\ &= y + \ln 2.\end{aligned}$$

Note that the MMAE estimate is smaller than the MMSE estimate since $\ln 2 \approx 0.6931$.

Finally, the mode of an exponentially distributed random variable with parameter λ is 0. What we have here is an exponential distribution with parameter $\lambda = 1$ shifted to the right by y. Hence,

$$\begin{aligned}\hat{\theta}_{MAP}(y) &= \mathrm{mode}[\Theta \,|\, Y = y] \\ &= y.\end{aligned}$$

Note that the MAP estimate is smaller than the MMAE estimate.

This is a nice example of a case where the MMSE, MMAE, and MAP estimates are all different.

6. 4 points. Kay I, problem 11.16.

Solution: This problem falls has a linear Gaussian model. So instead of re-deriving all of those results, we will just apply them here by noting that

$$Y = \begin{bmatrix} Y_{-M} \\ \vdots \\ Y_M \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & -M \\ \vdots & \vdots \\ 1 & M \end{bmatrix}}_{H} \underbrace{\begin{bmatrix} A \\ B \end{bmatrix}}_{\Theta} + \underbrace{\begin{bmatrix} W_{-M} \\ \vdots \\ W_M \end{bmatrix}}_{W}$$

To derive the MMSE estimator and its associated performance, we need to calculate

$$\begin{aligned} \mathrm{E}[\Theta \,|\, Y = y] &= \mu_\Theta + \Sigma_\Theta H^\top \left(H \Sigma_\Theta H^\top + \Sigma_W \right)^{-1} (y - H\mu_\Theta) \\ \mathrm{cov}[\Theta \,|\, Y = y] &= \Sigma_\Theta - \Sigma_\Theta H^\top \left(H \Sigma_\Theta H^\top + \Sigma_W \right)^{-1} H \Sigma_\Theta \end{aligned}$$

where it is given that

$$\begin{aligned} \mu_\Theta &= [A_0, B_0]^\top \\ \Sigma_\Theta &= \mathrm{diag}(\sigma_A^2, \sigma_B^2) \\ \Sigma_W &= \sigma^2 I. \end{aligned}$$

We've got everything we need, it is just a matter of doing the calculations. The only difficult part is the matrix inverse. If you just try to plug and chug here, you will see that you need to do a $2M+1 \times 2M+1$ matrix inverse. Instead, we can use the the matrix inversion lemma (sorry for the conflicting notation here)

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}.$$

with $A = \Sigma_W$, $B = H$, $C = H^\top$, and $D^{-1} = -\Sigma_\Theta$. Then we can write

$$\left(H \Sigma_\Theta H^\top + \Sigma_W \right)^{-1} = \frac{1}{\sigma^2} I - \frac{1}{\sigma^2} H \left(\Sigma_\Theta^{-1} + \frac{1}{\sigma^2} H^\top H \right)^{-1} H^\top \frac{1}{\sigma^2}.$$

The good news about this is that the new matrix inverse is just 2×2. Let's focus on that for a second

$$\begin{aligned} \left(\Sigma_\Theta^{-1} + \frac{1}{\sigma^2} H^\top H \right)^{-1} &= \left(\begin{bmatrix} \frac{1}{\sigma_A^2} & 0 \\ 0 & \frac{1}{\sigma_B^2} \end{bmatrix} + \begin{bmatrix} \frac{2M+1}{\sigma^2} & 0 \\ 0 & \frac{P}{\sigma^2} \end{bmatrix} \right)^{-1} \\ &= \begin{bmatrix} \frac{1}{\frac{1}{\sigma_A^2} + \frac{2M+1}{\sigma^2}} & 0 \\ 0 & \frac{1}{\frac{1}{\sigma_B^2} + \frac{P}{\sigma^2}} \end{bmatrix} \end{aligned}$$

where $P := \sum_{m=-M}^{M} m^2$. So plugging this result in and grinding out the rest of the linear algebra, we can write the MMSE/MMAE/MAP estimator as

$$\hat{\theta} = \begin{bmatrix} A_0 + \frac{\bar{y} - A_0}{\frac{\sigma^2}{(2M+1)\sigma_A^2} + 1} \\ B_0 + \frac{\left(\frac{1}{P} \sum_{m=-M}^{M} m y_m \right) - B_0}{\frac{\sigma^2}{P \sigma_A^2} + 1} \end{bmatrix}$$

where $\bar{y} = \frac{1}{2M+1} \sum_{m=-M}^{M} y_m$. The MMSE can be similarly computed as

$$\text{MMSE} = \begin{bmatrix} \left(\frac{1}{\sigma_A^2} + \frac{2M+1}{\sigma^2} \right)^{-1} \\ \left(\frac{1}{\sigma_B^2} + \frac{P}{\sigma^2} \right)^{-1} \end{bmatrix}$$

Both MMSE's benefit from narrower priors (smaller σ_A^2 and σ_B^2), as you would expect. Note, however, that $P := \sum_{m=-M}^{M} m^2$ grows faster in M than $2M + 1$. Hence, the MMSE of the slope is going to zero more quickly than the MMSE of the intercept as more observations arrive. We can conclude then that the intercept (parameter A) benefits the most from prior knowledge since the prior on the slope is "washed away" more quickly by the observations than the prior on the intercept.