1(a) The MAP estimate of θ is:

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} \{ P(D|\theta) \, P(\theta) \}$$

Here $D = \{X_1, X_2, \cdots X_n\}$ i.i.d. Gaussian random veriables with mean θ (unknown) and veriance σ_0^2 (known). So we can write,

$$P(D|\theta) = \prod_{i=1}^{n} P(X_i|\theta)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(X_i-\theta)^2}{2\sigma_0^2}}$$

As θ itself is selected from a Normal distribution $\mathcal{N}(\mu, \sigma^2)$, so we can write,

$$P(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}$$

Taking log of $P(D|\theta)$

$$\ln P(D|\theta) = \ln \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(X_i-\theta)^2}{2\sigma_0^2}} \right)$$

$$= \sum_{i=1}^{n} \left\{ \ln \frac{1}{\sqrt{2\pi\sigma_0^2}} + \ln e^{-\frac{(X_i-\theta)^2}{2\sigma_0^2}} \right\}$$

$$= \ln \frac{1}{\sqrt{2\pi\sigma_0^2}} + \sum_{i=1}^{n} -\frac{(X_i-\theta)^2}{2\sigma_0^2}$$

Taking log of $P(\theta)$

$$\ln P(\theta) = \ln \frac{1}{\sqrt{2\pi\sigma^2}} + \ln e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}$$

$$= \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(\theta-\mu)}{2\sigma^2}$$

step-1 Take log

We can write the log of θMAP as

$$\ln \{P(D|\theta)P(\theta)\} = \ln P(D|\theta) + \ln P(\theta)$$

$$= \ln \frac{1}{\sqrt{2\pi\sigma_0^2}} - \sum_{i=1}^{n} \frac{(X_i-\theta)^2}{2\sigma_0^2} + \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(\theta-\mu)^2}{2\sigma^2}$$

step-2 Drop constant

As σ_0, σ and μ are known we can write our objective function as

$$\underset{\theta}{\operatorname{argmax}} -\left\{\sum_{i=1}^{n} (X_i-\theta)^2 + (\theta-\mu)^2\right\}$$

we can minimize the negative likelihood function and formulate as:

$$\ln \theta_{MAP} = \underset{\theta}{\operatorname{argmin}} \left\{\sum_{i=1}^{n} (X_i-\theta)^2 + (\theta-\mu)^2\right\}$$

step-3 Taking derivative w.r.t. θ

$$\frac{d}{d\theta}(\theta_{MAP}) = 2 * \sum (X_i-\theta)(-1) + 2*(\theta-\mu)$$

Equating the derivative to zero (for stationary point):

$$-2\sum_{i=1}^{n}(X_i-\theta) + 2(\theta-\mu) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(X_i-\theta) = \theta - \mu$$

$$\Rightarrow \mu + \sum_{i=1}^{n} X_i = n\theta + \theta$$

$$\Rightarrow \theta_{MAP} = \frac{\mu + \sum_{i=1}^{n} X_i}{n+1}$$

which is the MAP estimate of θ.

1(b) Now θ is selected from a Laplas distribution. So we can rewrite the prior

$$P(\theta) = \frac{1}{2b} e^{\left(\frac{-|\theta-\mu|}{b}\right)}$$

where, μ is the mean of distribution (known)

and b is diversity (known).

We can write the likelihood function as

$$\theta_{MAP} = \arg\max_{\theta} \{ P(D|\theta)\, P(\theta) \}$$

Step-1 Taking the log of the MAP estimate.

$$\underset{\theta}{\text{argmax}} \ln \left\{ P(D|\theta)\, P(\theta) \right\} = \underset{\theta}{\text{argmax}} \ln \left\{ P(D|\theta)\, P(\theta) \right\}$$

$$= \underset{\theta}{\text{argmax}} \ln \left\{ \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{\frac{-(X_i-\theta)^2}{2\sigma_0^2}} * \frac{1}{2b} e^{\left(\frac{-|\theta-\mu|}{b}\right)} \right\}$$

$$= \underset{\theta}{\text{argmax}} \left\{ \sum_{i=1}^{n} \left\{ \ln\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right) + \ln e^{-\frac{(X_i-\theta)^2}{2\sigma_0^2}} \right\} + \ln \left\{ \frac{1}{2b} e^{-\frac{|\theta-\mu|}{b}} \right\} \right\}$$

$$= \underset{\theta}{\text{argmax}} \left\{ \ln \frac{1}{\sqrt{2\pi\sigma_0^2}} + \sum_{i=1}^{n} -\frac{(X_i-\theta)^2}{2\sigma_0^2} + \ln \frac{1}{2b} + \left(\frac{-|\theta-\mu|}{b}\right) \right\}$$

Step-2 Droping constent terms we can formulate the objective function as:

$$\underset{\theta}{\text{argmax}} - \left\{ \sum_{i=1}^{n} (X_i-\theta)^2 + |\theta-\mu| \right\}$$

minimize the negative log likelihood, we can write

$$\underset{\theta}{\text{argmin}} \left\{ \sum_{i=1}^{n} (X_i-\theta)^2 + |\theta-\mu| \right\}$$

Step-3 Now we take derivative w.r.t. θ and equating to zero we get

$$\frac{d}{d\theta}\sum_{i=1}^{n}(X_i-\theta)^2 + \frac{d}{d\theta}|\theta-\mu| = 0$$

$$\Rightarrow -2\sum_{i=1}^{n}(X_i-\theta) + \frac{d}{d\theta}|\theta| = 0 \left[\text{Assuming } \mu = 0\right]$$

Here the second term

$$\frac{d}{d\theta}|\theta| = \frac{\theta}{|\theta|} = \begin{cases} -1 & \text{if } \theta < 0 \\ +1 & \text{if } \theta > 0 \end{cases}$$

But not diferentiable at $\theta = 0$. So we can not get a closed form solution for this estimate.

One alternative for such a non smoth objective is to use proximal methods. Use Gradient descent for smooth component of the optimization $\left(\sum_{i=1}^{n}(X_i-\theta)^2\right)$ and then for the values of θ that are close to zero, set them to zero. (θ_t is the value of θ at tth iteration).

$$\text{prox}_{\eta\lambda l_1}(\theta_t) = \begin{cases} \theta_{t-1} - \eta\lambda & \text{if } \theta_{t-1} > \eta\lambda \\ 0 & \text{if } |\theta_{t-1}| < \eta\lambda \\ \theta_{t-1} + \eta\lambda & \text{if } \theta_{t-1} < -\eta\lambda \end{cases}$$

1(c) The MAP estimate of θ is,

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} \left\{ P(D|\theta)\, P(\theta) \right\}$$

$D = \{X_1, X_2, \ldots X_n\}$ are multivariate i.i.d Gaussian random variables with mean θ (known) and covarience Σ_0 (known) where $\theta \in \mathbb{R}^d$ and $\Sigma_0 = I \in \mathbb{R}^{d \times d}$ (I is the identity matrix).

$\theta \in \mathbb{R}^d$ itself is selected from a zero mean multivariate Gaussian $N(\mu = 0, \Sigma = \sigma^2 I)$ with known varience parameter σ^2 on a diagonal.

We can write,

$$P(D|\theta) = \prod_{i=1}^{n} P(X_i|\theta)$$

$$= \prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2} |\Sigma_0|^{1/2}} e^{-(1/2) \cdot (X_i - \theta)^T \Sigma_0^{-1} (X_i - \theta)}$$

$$= \prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2} |I|^{1/2}} e^{-(1/2) (X_i - \theta)^T (I)^{-1} (X_i - \theta)} \qquad \left[\Sigma_0^{d \times d} = I \right]$$

As $I^{-1} = I$ and $|I| = 1$. we can write.

$$P(D|\theta) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2}} e^{-(1/2)(X_i - \theta)^T (X_i - \theta)}$$

log of $P(D|\theta)$

$$\ln P(D|\theta) = \sum_{i=1}^{n} \left\{ -\frac{d}{2} \ln(2\pi) - \frac{1}{2}(X_i - \theta)^T (X_i - \theta) \right\}$$

Now,

$$P(\theta) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-(1/2)\theta^T \theta \Sigma^{-1}}$$

$$= \frac{1}{(2\pi)^{d/2} |\sigma^2 I|^{1/2}} e^{-(1/2)\theta^T\theta \cdot (\sigma^2 I)^{-1}}$$

As. $\underset{d\times d}{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & \cdots & 0 \\ - & - & - & - & - \\ 0 & 0 & - & \cdots & 1 \end{bmatrix}$ and $\underset{d\times d}{\sigma^2 I} = \begin{bmatrix} \sigma_{1,1}^2 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma_{22}^2 & \cdots & \cdots & 0 \\ - & - & - & - & - \\ 0 & 0 & \cdots & \cdots & \sigma_{dd}^2 \end{bmatrix}$

We can write

$$P(\theta) = \frac{1}{(2\pi)^{d/2} (\sigma^2)^{d/2}} e^{-(1/2)\theta^T \theta (\sigma^2)^{-1} I^{-1}}$$

As $I^{-1} = I$ we can write. $\left[\because |\sigma^2 I|^{1/2} = ((\sigma^2)^d)^{1/2} = \sigma^d \right]$

$$P(\theta) = \frac{1}{(2\pi)^{d/2} \sigma^d} e^{-(1/2)\theta^T\theta\sigma^{-2}}$$

$$P(\theta) = \frac{1}{(\sqrt{2\pi\sigma^2})^d} e^{-\frac{\theta^T\theta}{2\sigma^2}}$$

$$\therefore \ln P(\theta) = -\frac{d}{2}\ln(2\pi\sigma^2) - \frac{\theta^T\theta}{2\sigma^2}$$

Now $\ln \theta_{MAP} = \ln P(D|\theta) + \ln P(\theta)$

$$= \sum_{i=1}^{n} \left\{ -\frac{d}{2}\ln(2\pi) - \frac{1}{2}(x_i - \theta)^T (x_i - \theta) \right\} - \frac{d}{2}\ln(2\pi\sigma^2) - \frac{\theta^T\theta}{2\sigma^2}$$

droping constant we can write the objective function as.

$$\operatorname*{argmax}_{\theta} \left\{ \sum_{i=1}^{n} - \left\{ \frac{1}{2}(x_i - \theta)^T (x_i - \theta) \right\} - \frac{\theta^T\theta}{2\sigma^2} \right\}$$

we can minimize the negative loglikdihood

$$\operatorname*{argmin}_{\theta} \left\{ \sum_{i=1}^{n} \left\{ \frac{1}{2}(x_i - \theta)^T (x_i - \theta) \right\} + \frac{\theta^T\theta}{2\sigma^2} \right\}$$

we can further write

$$\operatorname*{argmin}_{\theta} \left\{ \sum_{i=1}^{n} \left\{ (x_i - \theta)^T \cdot (x_i - \theta) \right\} + \frac{\theta^T\theta}{\sigma^2} \right\}$$

Taking the partial derivative w.r.t θ_j $\{j = 1 \cdots d\}$

$$\frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} (X_i - \theta)^2 + \frac{\partial}{\partial \theta_j} \frac{(\theta)^2}{\sigma^2} = 0$$

Now

$$\frac{\partial}{\partial \theta_j} (X - \theta)^2 = \frac{\partial \sum_{i=1}^{d} (X_i - \theta_i)^2}{\partial \theta_j} = \frac{\partial \{(X_1 - \theta_1)^2 + \cdots (X_j - \theta_j)^2 + \cdots (X_d - \theta_d)^2\}}{\partial \theta_j}$$

$$\Rightarrow \frac{\partial (X_j - \theta_j)^2}{\partial \theta_j}$$

$$\Rightarrow -2(x_j - \theta_j)$$

and

$$\frac{\partial}{\partial \theta_j} (\theta)^2 = \frac{\partial \sum_{i=1}^{n} \theta_i^2}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \left(\frac{\theta_1^2 + \cdots \theta_j^2 + \cdots \theta_d^2}{\sigma^2} \right)$$

$$= \frac{\partial \theta_j^2}{\partial \theta_j} = 2\theta_j / \sigma^2$$

$$\therefore \frac{\partial}{\partial \theta_j} (X - \theta)^2 + \frac{\theta^2}{\sigma^2} = -2(x_j - \theta_j) + 2\theta_j/\sigma^2 = -2$$

$$\therefore \sum_{j=1}^{d} \frac{\partial}{\partial \theta_j} \left(\sum_{i=1}^{n} (X_i - \theta)^2 + \frac{\theta^2}{\sigma^2} \right) = 0$$

$$\Rightarrow \sum_{j=1}^{d} \frac{2\theta_j}{\sigma^2} + \sum_{j=1}^{d} \sum_{i=1}^{n} (-2)(x_{ij} - \theta_j) = 0$$

$$\Rightarrow \sum_{j=1}^{d} \frac{\theta_j}{\sigma^2} + \sum_{j=1}^{d} \sum_{i=1}^{n} (x_{ij} - \theta_j) = 0$$

$$\Rightarrow \overset{d \times 1}{\begin{bmatrix} \theta_1/\sigma^2 \\ \theta_2/\sigma^2 \\ \vdots \\ \theta_d/\sigma^2 \end{bmatrix}} + \overset{d \times 1}{\begin{bmatrix} \sum_{i=1}^{n} x_{i1} - \theta_1 \\ \sum_{i=1}^{n} x_{i2} - \theta_2 \\ \vdots \\ \sum_{i=1}^{n} x_{id} - \theta_d \end{bmatrix}} = 0$$

$$\Rightarrow \begin{bmatrix} \theta_1/\sigma^2 \\ \vdots \\ \theta_d/\sigma^2 \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{n} x_{i1} \\ \vdots \\ \sum_{i=1}^{n} x_{id} \end{bmatrix} - n \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} = 0$$

$$\Rightarrow \overset{d \times 1}{\theta}/\sigma^2 - n \overset{d \times 1}{\theta} + \sum_{i=1}^{n} \overset{d \times 1}{X_i} = 0$$

$$\Rightarrow \overset{d \times 1}{\theta} \left(-1/\sigma^2 + n\right) = \sum_{i=1}^{n} \overset{d \times 1}{X_i}$$

$$\Rightarrow \overset{d \times 1}{\theta_{MAP}} = \frac{\sum_{i=1}^{n} \overset{d \times 1}{X_i}}{n - \frac{1}{\sigma^2}}$$

Bonus Question (C)

The momentum v_t without bias correction is given as:

$$v_t = (1-\beta_2) \sum_{i=1}^{t} \beta_2^{t-i} \cdot g_i^2$$

The expected value of momentum (the exponential moving average over squared gradient) at time t can be calculated as:

$$E[v_t] = E\left[(1-\beta_2) \sum_{i=1}^{t} \beta_2^{t-i} \cdot g_i^2\right]$$

$$= E[g_t^2] \cdot (1-\beta_2) \sum_{i=1}^{t} \beta_2^{t-i} + \zeta$$

$$= E[g_t^2] \cdot (1-\beta_2^t) + \zeta$$

As $\zeta = 0$ for stationary $E[g_t^2]$ and should be chosen small. So we can write.

$$\therefore \quad E[v_t] = E[g_t^2] \cdot \underbrace{(1-\beta_2^t)}_{\text{Bias}}$$

Which is a biased estimate of expected value of squired gradiant. The bias term $(1-\beta_2^t)$ is caused by initializing the moving average with zero. Thats why the bias is corrected on later steps by dividing v_t by $(1-\beta_2^t)$

$$\hat{v}_t \leftarrow v_t / (1-\beta_2^t)$$

Similarly the expected value of momentum m_t (the exponential moving avarage over gradient) at time t can be written as.

$$E[m_t] = E\left[(1-\beta_1) \sum_{i=1}^{t} \beta_1^{t-i} \cdot g_i\right]$$

$$= E[g_t] \cdot (1-\beta_1) \sum_{i=1}^{t} \beta_1^{t-i} + \zeta$$

$$= E[g_t] \cdot (1-\beta_1^t) + \zeta$$

As ζ is small,

$$E[m_t] = E[g_t] \cdot \underbrace{(1-\beta_1^t)}_{\text{Bias}}$$

which is a biased estimate of gradient's expected value at time t.