1. Likelihood of point n: $\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$

$\therefore$ Likelihood of full data set:

$$p(x_1, x_2, \dots x_N | \mu, \sigma^2)$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Log likelihood is given by:

$$\log p(x_1, x_2, \dots x_N | \mu, \sigma^2) = -\sum_{i=1}^{N} \left\{ \frac{(x_i-\mu)^2}{2\sigma^2} + \log(\sqrt{2\pi}) + \log\sigma \right\}$$

$$= \frac{-1}{2\sigma^2} \sum_{i=1}^{N} (x_i-\mu)^2 - \frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2$$

$$\therefore \frac{d \log p()}{d\mu} = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i-\mu) = \frac{1}{\sigma^2} \left\{ \sum_{i=1}^{N} x_i - N\mu \right\}$$

Setting this to 0 gives the ML estimator:

$$\sum_{i=1}^{N} x_i - N\mu = 0 \quad \Rightarrow \quad \boxed{\mu_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i}$$

$$\frac{d \log p()}{d\sigma^2} = \frac{1}{2\sigma^4} \sum_{i=1}^{N} (x_i-\mu)^2 - \frac{N}{2\sigma^2}$$

$$= \frac{1}{2\sigma^2} \left(\frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i-\mu)^2 - N \right)$$

Setting this to 0 gives the ML estimator:

$$\boxed{\sigma^2_{ML} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_{ML})^2}$$

Expected values:

$$E[\mu_{ML}] = \frac{1}{N} E\left[\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} E(x_i)$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mu \qquad (\text{Data from } \mathcal{N}(\mu, \sigma^2))$$

$$= \underline{\mu}$$

$$E[\sigma^2_{ML}] = \frac{1}{N} E\left[\sum_{i=1}^{N} (x_i - \mu_{ML})^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} E\left[(x_i - \mu_{ML})^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} E\left[x_i^2 + \mu_{ML}^2 - 2x_i\mu_{ML}\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} \left\{E[x_i^2] + E[\mu_{ML}^2] - 2E[x_i\mu_{ML}]\right\}$$

Substitute the value of μ_{ML} from earlier:

$$= \frac{1}{N}\sum_{i=1}^{N} \left\{E[x_i^2] + E\left[\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} x_j x_k\right] - \frac{2}{N} E\left[x_i \sum_{j=1}^{N} x_j\right]\right\}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \left\{E[x_i^2] + \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} E[x_j x_k] - \frac{2}{N}\sum_{j=1}^{N} E[x_i x_j]\right\}$$

Now we need the result: $E[x_n x_m] = \mu^2 + I_{nm}\sigma^2$

(See Bishop, Exercise 1.12 - solved online)

Thus we get:

$$E[\sigma^2_{ML}] = \frac{1}{N}\sum_{i=1}^{N}\left\{E[x_i^2] + \frac{1}{N^2}(\mu^2N^2 + \sigma^2 N) - \frac{2}{N}(\mu^2 N + \sigma^2)\right\}$$

Also, $E[x_i^2] - (E[x_i])^2 = \sigma^2$, by defn.

$$\Rightarrow E[x_i^2] = \sigma^2 + \mu^2$$

$$\Rightarrow E[\sigma^2_{ML}] = \frac{1}{N}\sum_{i=1}^{N}\left\{\sigma^2 \cancel{+\mu^2} \cancel{+\mu^2} + \frac{\sigma^2}{N} \cancel{-2\mu^2} - \frac{2\sigma^2}{N}\right\}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left\{\sigma^2\left(1-\frac{1}{N}\right)\right\}$$

$$= \underline{\underline{\left(\frac{N-1}{N}\right)\sigma^2}}$$

2. Likelihood:

$$p\left(\tfrac{M}{N}\text{ heads} \mid p_H\right) = p_H^M (1-p_H)^{N-M}$$

Log likelihood:

$$\log p() = M \log p_H + (N-M)\log(1-p_H)$$

$$\frac{d \log p()}{d p_H} = \frac{M}{p_H} + \frac{M-N}{1-p_H}$$

Setting to 0, we get: $\dfrac{M}{p_H} = \dfrac{N-M}{1-p_H}$

$$\Rightarrow M - \cancel{M p_H} = N p_H - \cancel{M p_H} \Rightarrow \boxed{p_{H_{ML}} = \frac{M}{N}}$$

Bayesian approach:

$$p\left(p_H \,\Big|\, \tfrac{M}{N}\text{ heads}\right) = \frac{p\left(\tfrac{M}{N}\text{ heads} \,\big|\, p_H\right) \cdot p(p_H)}{p\left(\tfrac{M}{N}\text{ heads}\right)}$$

$$\sim p_H^{M}(1-p_H)^{N-M} \cdot 6\,p_H(1-p_H)$$

$$\sim p_H^{M+1}(1-p_H)^{N-M+1}$$

$\therefore$ By analogy with above, $\boxed{p_{H_{MAP}} = \dfrac{M+1}{N\cancel{-M}+2}}$

3. $$\Phi \underbrace{(\Phi^T\Phi)^{-1}\Phi^T \underline{t}}_{\text{Define } = \, \underline{\tilde{t}} \, = \, (\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_M)^T} = \underline{y} \quad \text{(the least-squares soln.)}$$

$$\underline{y} = \Phi \underline{\tilde{t}} = \underline{\Psi}_1 \tilde{t}_1 + \underline{\Psi}_2 \tilde{t}_2 + \ldots + \underline{\Psi}_M \tilde{t}_M$$

where $\underline{\Psi}_1, \underline{\Psi}_2, \ldots \underline{\Psi}_M$ are the columns of Φ

To see that this is an orthogonal projection:

For any $\underline{\Psi}_i$ we have:

$$\circledast \quad \Phi(\Phi^T\Phi)^{-1}\Phi^T\underline{\Psi}_i = [\Phi\cancel{(\Phi^T\Phi)^{-1}}\,\cancel{\Phi^T\Phi}]_i = \underline{\Psi}_i$$

Thus:

$$(\underline{y} - \underline{t})^T \underline{\Psi}_i = (\Phi(\Phi^T\Phi)^{-1}\Phi^T\underline{t} - \underline{t})^T \underline{\Psi}_i$$

$$= \underline{t}^T(\Phi(\Phi^T\Phi)^{-1}\Phi^T - I)^T \underline{\Psi}_i$$

$$= \underline{t}^T(\underline{\Psi}_i^T\Phi(\Phi^T\Phi)^{-1}\Phi^T - \underline{\Psi}_i^T)^T$$

$$= \underline{t}^T(\underline{\Psi}_i^T - \underline{\Psi}_i^T)^T \quad (\text{from } \circledast)$$

$$= \underline{\underline{0}}$$ $\therefore$ $\underline{y} - \underline{t}$ orthogonal to all columns of Φ.

4. Suppose C_j denotes ~~the class with the highest posterior~~

The loss expected in assigning data point $\underline{x}$ to class C_j is:

$$L_E(j) = \sum_{i=1}^{K} P(C_i|\underline{x}) . L_{ij}$$

$\therefore$ our optimal decision criterion is as follows:

Let $k = \underset{j}{\operatorname{argmin}}\, L_E(j)$

If $L_E(k) < \lambda$, assign class C_k

Else, reject.

Now suppose L is given by $1_{K\times K} - I_{K\times K}$

In this case the expected loss becomes:

$$L_E(j) = \sum_{i=1}^{K} P(C_i|\underline{x}) L_{ij} = \sum_{i \neq j} P(C_i|\underline{x})$$

$$= \underline{1 - P(C_j|\underline{x})}$$

So $k = \underset{j}{\operatorname{argmin}}\, L_E(j)$ gives the class C_k with the highest posterior

Thus we simply have:

If $L_E(k) = 1 - P(C_k|\underline{x}) < \lambda$, assign C_k

$$\Rightarrow \underline{P(C_k|\underline{x}) > 1 - \lambda}$$

is the acceptance criterion; which corresponds to a rejection threshold of $\boxed{\theta = 1 - \lambda}$