Part 1:

1. x_n and x_m are two different point sampled from a Gaussian distribution with mean μ and variance σ^2,

it means $\int_{-\infty}^{\infty} N(x \mid \mu, \sigma^2)\, dx = 1$

so, $E[x_n] = \int_{-\infty}^{\infty} N(x_n \mid \mu, \sigma^2)\, x_n\, dx_n = \mu$

$E[x_m] = \int_{-\infty}^{\infty} N(x_m \mid \mu, \sigma^2)\, x_m\, dx_m = \mu$

if $x_n \neq x_m$, $E[x_n \cdot x_m] = E[x_n] \cdot E[x_m] = \mu^2$

elif $x_n = x_m$, $E[x_n \cdot x_m] = E[x^2] = \int_{-\infty}^{\infty} N(x \mid \mu, \sigma^2)\, x^2\, dx$

$= \mu^2 + \sigma^2$

so, $E[x_n x_m] = \mu^2 + I_{nm}\, \sigma^2$,

$$I_{nm} = \begin{cases} 1, & \text{if } n = m \\ 0, & \text{elif } n \neq m \end{cases} \quad ※$$

$$L(\mu, \sigma^2 \mid X) = p(X \mid \mu, \sigma^2) = \prod_{n=1}^{N} N(x_n \mid \mu, \sigma^2)$$

取 log, $\ln p(X \mid \mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi)$

Maximizing with respect to μ,

can get $\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n$, $E[\mu_{ML}] = \frac{1}{N} \sum E[x_n]$

$= \frac{1}{N} (E[x_1] + E[x_2] + \dots + E[x_N])$

$= \frac{1}{N} \cdot (N \cdot \mu) = \mu$ —— (4)

Maximizing with respect to σ^2,

can get $\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})^2$,

$$E[\sigma_{ML}^2] = \frac{1}{N} \sum E[(x_n - \mu_{ML})^2] = \frac{1}{N} \sum \left(\underline{E[x_n^2]} - \underline{2E[x_n \mu_{ML}]} + \underline{E[\mu_{ML}^2]} \right)$$

$$= \underline{\frac{1}{N} \cdot N (\mu^2 + \sigma^2)} - \underline{2E[x_n \cdot \frac{1}{N} \sum_{n=1}^{N} x_n]} + \underline{E[\frac{1}{N^2} \sum_{n=1}^{N} x_n \sum_{n=1}^{N} x_n]}$$

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

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

2. a and b are two independent random vectors,

$p(a,b) = p(a)p(b)$

$$\vec{y} = \vec{a}+\vec{b} \Rightarrow y = \frac{\vec{a}+\vec{b}}{2} = \frac{\vec{a}}{2}+\frac{\vec{b}}{2} \quad *$$

$$\mathrm{cov}(\vec{a}+\vec{b}) = E[(\vec{a}+\vec{b}-E[\vec{a}+\vec{b}])(\vec{a}+\vec{b}-E[\vec{a}+\vec{b}])^T] = \mathrm{cov}(\vec{y})$$

$$\mathrm{cov}(\vec{a}) = E[(\vec{a}-E[\vec{a}])(\vec{a}-E[\vec{a}])^T]$$

$$\mathrm{cov}(\vec{b}) = E[(\vec{b}-E[\vec{b}])(\vec{b}-E[\vec{b}])^T]$$

$$\because E[\vec{a}+\vec{b}] = E[\vec{a}]+E[\vec{b}]$$

$$\mathrm{cov}(\vec{a})+\mathrm{cov}(\vec{b}) = E[(\vec{a}-E[\vec{a}]+\vec{b}-E[\vec{b}])(\vec{a}-E[\vec{a}]+\vec{b}-E[\vec{b}])^T]$$

$$= E[(\vec{a}+\vec{b}-E[\vec{a}+\vec{b}])(\vec{a}+\vec{b}-E[\vec{a}+\vec{b}])^T]$$

$$= \mathrm{cov}(\vec{a}+\vec{b}) = \mathrm{cov}(\vec{y}) \quad \#$$

3. $\sigma_N^2(x) = \frac{1}{\beta} + \phi(x)^T S_N \phi(x)$

使用反證法，$\sigma^2_{N+1}(x) \le \sigma^2_N(x)$

$$\Rightarrow \phi(x)^T \underline{S_{N+1}} \phi(x) \le \phi(x)^T S_N \phi(x)$$

use Appendix C,

$$S_{N+1} = [S_N^{-1} + \beta\phi(x)\phi(x)^T]^{-1} = S_N - \frac{S_N\phi(x)\phi(x)^T S_N}{1+\phi(x)^T S_N \phi(x)}$$

so, if $\phi(x)^T \frac{S_N\phi(x)\phi(x)^T S_N}{1+\phi(x)^T S_N\phi(x)} \phi(x) \ge 0$ 則成立，

$\Rightarrow \phi(x)^T S_N \phi(x) \ge 0$. as long as the covariance matrix S_N is positive semi-definite

4. A Gaussian noise ϵ_i is added to each input variable x_i,

So let
$$\tilde{y}_n = w_0 + \sum_{i=1}^{D} w_i (x_{ni} + \epsilon_{ni})$$
$$= y_n + \sum_{i=1}^{D} w_i \epsilon_{ni}$$

where $y_n = y(x_n, w)$, $\epsilon_{ni} \sim N(0, \delta^2)$

so,
$$\tilde{E} = \frac{1}{2} \sum \{ \tilde{y}_n - t_n \}^2$$
$$= \frac{1}{2} \sum \{ \tilde{y}_n^2 - 2\tilde{y}_n t_n + t_n^2 \}.$$
$$= \frac{1}{2} \sum \{ y_n^2 + \cancel{2 y_n \sum_{i=1}^{D} w_i \epsilon_{ni}} + \Big(\sum_{i=1}^{D} w_i \epsilon_{ni} \Big)^2$$
$$- 2 y_n t_n - \cancel{2 t_n \sum_{i=1}^{D} w_i \epsilon_{ni}} + t_n^2 \}$$

when taking the expectation of $\tilde{E}$,

since $E[\epsilon_i] = 0$, the second and fifth terms become zeros,

and for the third term, can get

$$E\Big[\Big(\sum_{i=1}^{D} w_i \epsilon_{ni} \Big)^2 \Big] = \sum_{i=1}^{D} w_i^2 \delta^2$$

so
$$E[\tilde{E}] = E[E_D] + \frac{1}{2} \sum_{i=1}^{D} w_i^2 \delta^2,$$

it shows that minimizing E_D with noise distribution is equivalent to minimizing E_D with noise-free input and a addition of a weight-decay regularization term.