3.3 Given the error function $E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\}^2$
To minimize it, we can take its derivative w.r.t $\mathbf{w}$ and set it to 0

$$\frac{dE_D(\mathbf{w})}{d\mathbf{w}} = \frac{d\left(\frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\}^2\right)}{d\mathbf{w}}$$

$$= \frac{1}{2}\sum_{n=1}^{N} r_n \frac{d\left(\{t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\}^2\right)}{d\mathbf{w}}$$

Using 3.13 from PRML,

$$= \sum_{n=1}^{N} r_n\{t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\}\boldsymbol{\phi}(\mathbf{x}_n)^T$$

Setting this gradient to zero gives:

$$0 = \sum_{n=1}^{N} r_n\{t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\}\boldsymbol{\phi}(\mathbf{x}_n)^T$$

$$\implies 0 = \sum_{n=1}^{N} \left(r_n t_n \boldsymbol{\phi}(\mathbf{x}_n)^T - r_n \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n)^T\right)$$

$$\implies 0 = \sum_{n=1}^{N} r_n t_n \boldsymbol{\phi}(\mathbf{x}_n)^T - \sum_{n=1}^{N} r_n \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n)^T$$

$$\implies 0 = \sum_{n=1}^{N} r_n t_n \boldsymbol{\phi}(\mathbf{x}_n)^T - \sum_{n=1}^{N} \mathbf{w}^T(\boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n)^T)r_n \qquad \text{since } r_n \text{ is a scalar}$$

$$\implies 0 = \sum_{n=1}^{N} r_n t_n \boldsymbol{\phi}(\mathbf{x}_n)^T - \mathbf{w}^T\sum_{n=1}^{N}(\boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n)^T)r_n$$

Let $\mathbf{R}$ be a diagonal matrix where the n^{th} diagonal value is given by r_n.

$$\implies 0 = (\boldsymbol{\Phi}^T\mathbf{R}\mathbf{t})^T - \mathbf{w}^T\boldsymbol{\Phi}^T\mathbf{R}\boldsymbol{\Phi}$$

$$\implies \mathbf{w}^T\boldsymbol{\Phi}^T\mathbf{R}\boldsymbol{\Phi} = (\boldsymbol{\Phi}^T\mathbf{R}\mathbf{t})^T$$

Taking transpose of both sides, we get:

$$\implies \mathbf{\Phi}^T\mathbf{R}\mathbf{\Phi}\mathbf{w} = (\mathbf{\Phi}^T\mathbf{R}\mathbf{t})$$

$$\implies \mathbf{w}_\star = (\mathbf{\Phi}^T\mathbf{R}\mathbf{\Phi})^{-1}(\mathbf{\Phi}^T\mathbf{R}\mathbf{t})$$

Another method for solving the same would be:

Let $\sqrt{r_n}\boldsymbol{\phi}(\mathbf{x}_n) = \boldsymbol{\phi}'(\mathbf{x}_n)$, and $\sqrt{r_n}t_n = t'_n$.

Then, the above expression becomes

$$0 = \sum_{n=1}^{N} t'_n\boldsymbol{\phi}'(\mathbf{x}_n)^T - \mathbf{w}^T\sum_{n=1}^{N}(\boldsymbol{\phi}'(\mathbf{x}_n)\boldsymbol{\phi}'(\mathbf{x}_n)^T)$$

Using 3.15 from PRML, we get a very similar result:

$$\mathbf{w}^\star = (\mathbf{\Phi'}^T\mathbf{\Phi'})^{-1}\mathbf{\Phi'}^T\mathbf{t}'$$

where $\mathbf{\Phi'} = \begin{bmatrix} \sqrt{r_1}\phi_0(\mathbf{x}_1) & \sqrt{r_1}\phi_1(\mathbf{x}_1) & \dots & \sqrt{r_1}\phi_{M-1}(\mathbf{x}_1) \\ \sqrt{r_2}\phi_0(\mathbf{x}_2) & \sqrt{r_2}\phi_1(\mathbf{x}_2) & \dots & \sqrt{r_2}\phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \dots & \vdots \\ \sqrt{r_N}\phi_0(\mathbf{x}_N) & \sqrt{r_N}\phi_1(\mathbf{x}_N) & \dots & \sqrt{r_N}\phi_{M-1}(\mathbf{x}_N) \end{bmatrix}$

and

$$\mathbf{t}' = \begin{bmatrix} \sqrt{r_1}t_1 \\ \sqrt{r_2}t_2 \\ \vdots \\ \sqrt{r_N}t_N \end{bmatrix}$$

Interpretations:

1. Data dependent noise variance : we assumed initially that the target variable t is given by $t_i = \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + \epsilon$ for all i. In this case, this equation becomes $\sqrt{r_i}t_i = \sqrt{r_i}\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + \epsilon \implies t_i = \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + \frac{\epsilon}{\sqrt{r_i}}$. Since ϵ is a zero mean Gaussian random variable that represents noise, we can see that the noise is having data dependent variance where $\sigma'_i = (\frac{1}{\sqrt{r_i}})^2\sigma = \frac{\sigma}{r_i}$.

2. Replicated data points : Data points with higher r_i value associated with them are replicated more in the dataset. This confirms the above result where higher r_i is associated with lower variance, as more data points that are replicated in the dataset imply reduced uncertainty around that data point.