Problem 3.2

So we first need to prove that matrix $\Psi = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top$ projects any N-dimensional vector v onto the subspace spanned by M columns of Φ (lets denote this subspace as $S(\Phi)$). Here we just assume $(\Phi^\top\Phi)^{-1}$ exists (i.e. $\Phi^\top\Phi$ is invertible) since it is a part of the definition of the matrix given in the problem condition.

Let us consider any N-dimensional vector v. We need to prove that there exist $\alpha_1 \ldots \alpha_M \in$ R such that $\Psi v = \alpha_1\varphi_1(D) + \ldots + \alpha_M\varphi_M(D)$. If we denote $\alpha = (\alpha_1, \ldots, \alpha_M)^\top$, then we need to prove there exists M-dimensional vector α such that $\Phi \cdot \alpha = \Psi v$. We notice now that $\alpha = (\Phi^\top\Phi)^{-1}\Phi^\top v$ is the very vector we are looking for, so it exists, so we proved that Ψ indeed projects v onto the subspace of columns of Φ.

Lets us now consider $w_{ML} = (\Phi^\top\Phi)^{-1}\Phi^\top t$. We need to prove $y = \Phi w_{ML}$ is an orthogonal projection of t onto the subspace of columns of Φ. This means, we need to prove $y - t \perp S(\Phi)$. This is the same as proving that $\Phi(\Phi^\top\Phi)^{-1}\Phi^\top t - t \perp S(\Phi)$.

Consider left part of the statement and multiply it by $\Phi^\top$. This gives us $\Phi^\top(\Phi(\Phi^\top\Phi)^{-1}\Phi^\top t - t) = (\Phi^\top\Phi)(\Phi^\top\Phi)^{-1}\Phi^\top t - \Phi^\top t = 0$. So, we see that all the columns of Φ are orthogonal with $\Phi(\Phi^\top\Phi)^{-1}\Phi^\top t - t$, which means $\Phi(\Phi^\top\Phi)^{-1}\Phi^\top t$ is an orthogonal projection of t onto $S(\Phi)$.

Problem 3.3

Since w^* is extremum, we can equate E_D gradient to zero:

$$\nabla E_D = -\sum_{n=1}^{N} r_n(t_n - w^\top\varphi(x_n))\varphi^\top(x_n) = 0 \tag{1}$$

Let us consider $R = \begin{pmatrix} r_1 & 0 & ... & 0 \\ 0 & r_2 & ... & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & ... & r_n \end{pmatrix}$

We can rewrite equation 1 in the following way:

$$\nabla E_D = -\Phi^\top Rt + \Phi^\top R\Phi w = 0 \tag{2}$$

Thus we obtain

$$w^* = (\Phi^\top R\Phi)^{-1}\Phi^\top Rt \tag{3}$$

We can consider the matrix R, one the one hand, as inverse data-dependent noise variance: different x_i will have different correspondent r_i, and the smaller r_i is, the smaller is the impact of i_{th} sample. So, r_i can be used as our confidence in the t_i value.

On the other hand, at least when r_i is integer, it can be considered as the number of times sample (x_i, t_i) was present in the dataset.

Problem 3.4

Let us average error function over all possible noise values, i.e. let us compute it's expected value with respect to added noise $\{\epsilon_i\}$. Lets us denote $x'_n = x_n + \epsilon_n$ - input variable with added noise.

$$\mathrm{E}[E_D] = \mathrm{E}[\frac{1}{2}\sum_{n=1}^{N}\{y(x'_n, w) - t_n\}^2] = \tag{4}$$

$$= \mathrm{E}[\frac{1}{2}\sum_{n=1}^{N}\{w_0 + \sum_{i=1}^{D} w_i(x_{ni} + \epsilon_{ni}) - t_n\}^2 = \tag{5}$$

$$= \mathrm{E}[\frac{1}{2}\sum_{n=1}^{N}\{w_0 + \sum_{i=1}^{D} w_i(x_{ni} + \epsilon_{ni}) - t_n\} = \tag{6}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\mathrm{E}[\{w_0 - t_n + \sum_{i=1}^{D} w_i(x_{ni} + \epsilon_{ni})\}^2] = \tag{7}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\mathrm{E}[(w_0 - t_n)^2 + (w_0 - t_n)\sum_{i=1}^{D} w_i(x_{ni} + \epsilon_{ni}) + \sum_{i,j=1}^{D} w_i w_j (x_{ni} + \epsilon_{ni})(x_{nj} + \epsilon_{nj})] = \tag{8}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\{(w_0 - t_n)^2 + (w_0 - t_n)\sum_{i=1}^{D} w_i x_{ni} + \sum_{i,j=1}^{D} w_i w_j (x_{ni}x_{nj} + \delta_{ij}\sigma^2)\} = \tag{9}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2 + \frac{N\sigma^2}{2} w^\top w \tag{10}$$

So, as expected, we see that error function, averaged over noise values, gives us weight-decayed sum-of-squares error function over noise-free input variables with omitted bias in regularization term, so minimizing the latter gives the same result as minimizing the former.