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.

Problem 3.5

Suppose we want to minimize $E_D(w)$ subject to $\sum\limits_{j=1}^{M} |w_j|^q \leq \eta$ This is equivalent to minimizing Lagrange function $L(w, \lambda) = E_D(w) + \lambda'(\sum\limits_{j=1}^{M}(|w_j|^q - \eta)$ under conditions that $\lambda' \geq 0$, $\sum\limits_{j=1}^{M} |w_j|^q \leq \eta$ and $\lambda'(\sum\limits_{j=1}^{M} |w_j|^q - \eta) = 0$. When we use regularization, we suppose we only vary $\{w_j\}$ while keeping λ fixed. So we can substitute $\lambda' = \frac{\lambda}{2}$ and pay no attention to η since it is constant, and try to minimize the Lagrangian which now takes form of $E_D(w) + \frac{\lambda}{2}(\sum\limits_{j=1}^{M}(|w_j|^q)$, and this is exactly the regularized least squares error function.

To find the dependence between η and w we note that for optimal solution $\{w_j^*\}$ we have $\eta = \sum\limits_{j=1}^{M}(|w_j^*|^q)$.

As of dependence between η and λ we can see that the greater λ is, the more is regularization effect and so the less will be η.

Problem 3.6

This problem looks simple, but in fact to understand it deeply one need to perform certain actions. First of all, we will write down the log-likelihood of the dataset with given W:

$$L(D) = \prod_{n=1}^{N}(2\pi)^{-\frac{k}{2}}|\Sigma|^{-\frac{1}{2}}\exp(-\frac{1}{2}(t_n - W^\top\varphi(X_n))^\top\Sigma^{-1}(t_n - W^\top\varphi(X_n)) \tag{11}$$

$$\log L(D) = -\frac{kN}{2}\log(2\pi) - \frac{N}{2}|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(t_n - W^\top\varphi(X_n))^\top\Sigma^{-1}(t_n - W^\top\varphi(X_n)) \tag{12}$$

We now want to find the maximum of the log-likelihood with respect to W to find W_{ML}. To achieve this, we will use matrix derivatives notation (see `https://en.wikipedia.org/wiki/Matrix_calculus`)

In fact, we see that taking derivative of scalar with respect to matrix gives us a matrix of the same size:

$$\frac{\partial \log L(D)}{\partial W} = \begin{pmatrix} \frac{\partial \log L(D)}{\partial W_{11}} & \frac{\partial \log L(D)}{\partial W_{12}} & \dots & \frac{\partial \log L(D)}{\partial W_{1K}} \\ \frac{\partial \log L(D)}{\partial W_{21}} & \frac{\partial \log L(D)}{\partial W_{22}} & \dots & \frac{\partial \log L(D)}{\partial W_{2K}} \\ \dots & \dots & \dots & \dots \\ \frac{\partial \log L(D)}{\partial W_{M1}} & \frac{\partial \log L(D)}{\partial W_{M2}} & \dots & \frac{\partial \log L(D)}{\partial W_{MK}} \end{pmatrix}$$

To get the work done, we first prove following identity:

$$\frac{\partial (Xa+b)^\top C(Xa+b)}{\partial X} = (C + C^\top)(Xa+b)a^\top \tag{13}$$

Here, C is a square matrix of some size $N \times N$, b is a vector of size N, a is a vector of size M, and X is $M \times N$ matrix.

To prove it, let's use write down the numerator in non-matrix form:

$$(Xa+b)^\top C(Xa+b) = \sum_{l=1}^{N}(\sum_{k=1}^{N}(\sum_{\beta=1}^{M} X_{l\beta}a_\beta + b_l)C_{kl})(\sum_{\gamma=1}^{N} X_{k\gamma}a_\gamma + b_k) \tag{14}$$

Taking now derivative of this with respect to X_{ij} we get:

$$\begin{aligned} \frac{\partial (Xa+b)^\top C(Xa+b)}{\partial X_{ij}} &= \sum_{k=1}^{N} a_j C_{ki}(X_k a + b_k) + \sum_{l=1}^{N}(X_l a + b_l)C_{il}a_j = && (15) \\ &= a_j[{C^i}^\top(Xa+b) + (Xa+b)^\top C_i] = && (16) \\ &= (C + C^\top)_i(Xa+b)a_j && (17) \end{aligned}$$

So, we see that $\frac{\partial (Xa+b)^\top C(Xa+b)}{\partial X} = (C + C^\top)(Xa+b)a^\top$.

Now, returning to the original problem, we have

$$\begin{aligned} \frac{\partial \log L(D)}{\partial W} &= -\frac{1}{2}\sum_{n=1}^{N}(\Sigma^{-1} + {\Sigma^{-1}}^\top)(t_n - W^\top\varphi(X_n))\varphi(X_n)^\top = && (18) \\ &= \Sigma^{-1}(T^\top - W^\top\Phi^\top)\Phi = 0 && (19) \end{aligned}$$

Just a reminder, T is a $N \times K$ matrix, $\Phi(X)$ is a $N \times M$ matrix, and W is $M \times K$ matrix.

So assigning this to 0 and reducing Σ^{-1} we get $T^\top\Phi = W_{ML}^\top\Phi^\top\Phi$ and so $W_{ML} = (\Phi^\top\Phi)^{-1}\Phi^\top T$ which shows us that indeed i-th column of W_{ML} has the same well known form of $W_{ML_i} = (\Phi^\top\Phi)^{-1}\Phi^\top t_i$ and W_{ML} is independent of Σ.

Now let us consider

$$\begin{aligned} \frac{\partial \log L(D)}{\partial \Sigma^{-1}} &= -\frac{N}{2}\frac{\partial \log|\Sigma|}{\partial \Sigma^{-1}} - \frac{1}{2}\sum_{n=1}^{N}(t_n - W_{ML}^\top\varphi(X_n))(t_n - W_{ML}^\top\varphi(X_n))^\top = && (20) \\ &= \frac{N}{2}\Sigma - \frac{1}{2}(T^\top - W^\top\Phi^\top)(T^\top - W^\top\Phi^\top)^\top && (21) \end{aligned}$$

So we obtain $\Sigma = \frac{1}{N} \sum_{n=1}^{N} (t_n - W_{ML}^{\top} \varphi(X_n))(t_n - W_{ML}^{\top} \varphi(X_n))^{\top}$.