Chapter 3 - Questions and Solutions

Edwin Fennell

3.1 Prove the least squares optimal solution for the linear regression case given in Eq. (3.13).

We seek to find the value of $\boldsymbol{\theta}$ that minimises

$$\sum_{n=1}^{N}(y_n - \boldsymbol{\theta}^T \mathbf{x}_n)^2$$

This is a differentiable function w.r.t all components of $\boldsymbol{\theta}$. Moreover, note that it is quadratic with a positive quadratic coefficient for all components of $\boldsymbol{\theta}$, guaranteeing that the sole stationary point is a global minimum.

Taking the derivative w.r.t θ_i gives

$$\sum_{n=1}^{N}(\mathbf{x}_n)_i \left(\left(\sum_{j=1}^{N} 2(\mathbf{x}_n)_j \theta_j \right) - 2y_n \right)$$

For the value $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ for which this is 0 for all i, we have

$$\sum_{n=1}^{N} \hat{\boldsymbol{\theta}}^T \mathbf{x}_n (\mathbf{x}_n)_i = \sum_{n=1}^{N} y_n (\mathbf{x}_n)_i$$

The right-hand side is the ith element of

$$\sum_{n=1}^{N} y_n \mathbf{x}_n$$

while the left-hand side is the ith element of

$$\left(\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T \right) \hat{\boldsymbol{\theta}}$$

Therefore these two vectors are equal and we have our requried equality.

3.2 Let $\hat{\theta}_i,\ i = 1, 2, ..., m$ be unbiased estimators of a parameter vector θ, so that $\mathbb{E}[\hat{\theta}_i] = \theta,\ i = 1, 2, ..., m$. Moreover, assume that the respective estimators are uncorrelated to each other and that all have the same variance $\sigma^2 = \mathbb{E}[(\theta_i - \theta)^T(\theta_i - \theta)]$. Show that by averaging the estimates, e.g.

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i$$

the new estimator has total variance $\sigma_c^2 = \mathbb{E}[(\theta_i - \theta)^T(\theta_i - \theta)] = \frac{\sigma^2}{m}$.

Trivially, the mean of our unbiased estimators is also an unbiased estimator. The variance of our estimator $\hat{\theta}$ is therefore

$$\mathbb{E}(\hat{\theta}^T\hat{\theta}) - \mathbb{E}(\theta)^T\mathbb{E}(\theta)$$

This expands to

$$\mathbb{E}\left(\sum_{i=1}^{m}\frac{1}{m}\hat{\theta_j}^T\left(\sum_{j=1}^{m}\frac{1}{m}\hat{\theta_j}\right)\right) - \mathbb{E}(\theta)^T\mathbb{E}(\theta)$$

The estimators are all pairwise uncorrelated, which means the the product of the expectations of any two distinct estimators is equal to the expectation of their product. Therefore we can rewrite the above as

$$\sum_{i=1}^{m}\left(\sum_{j=1}^{m}\frac{1}{m^2}\mathbb{E}(\hat{\theta}_i)^T\mathbb{E}(\hat{\theta}_j) + \frac{1}{m^2}(\mathbb{E}(\hat{\theta}_i^T\hat{\theta}_i) - \mathbb{E}(\hat{\theta}_i)^T\mathbb{E}(\hat{\theta}_i))\right) - \mathbb{E}(\theta)^T\mathbb{E}(\theta)$$

The first term with the double sum is actually just

$$\left(\sum_{i=1}^{m}\frac{1}{m}\mathbb{E}(\hat{\theta}_i)\right)^T\left(\sum_{i=1}^{m}\frac{1}{m}\mathbb{E}(\hat{\theta}_i)\right) = E(\theta)^TE(\theta)$$

This just cancels out with the last term, and we are left with the middle term, which is just $\frac{1}{m^2}$ times the sum of the variances of the initial m estimators, as required.

3.3 Let x be random variable uniformly distributed on $[0, \frac{1}{\theta}], \theta > 0$. Assume that g is a Lebesgue measurable function on $[0, \frac{1}{\theta}]$. Show that if $\hat{\theta} = g(x)$ is an unbiased estimator, then

$$\int_0^{\frac{1}{\theta}} g(x)dx = 1$$

Assume that $\hat{\theta}$ is an unbiased estimator. Then regardless of the value of θ

$$\theta = \mathbb{E}(\hat{\theta}) = \mathbb{E}(g(x)) = \int_0^{\frac{1}{\theta}} g(x)\phi(x)dx$$

where ϕ is the p.d.f of x. Since x is uniform, the p.d.f is just constantly θ This directly gives

$$\int_0^{\frac{1}{\theta}} g(x)dx = 1$$

We note that there is no function $g(x)$ s.t. this holds for all θ. We note that our condition gives

$$\int_a^b g(x)dx = 0 \ \forall (0 < a < b)$$

Therefore

$$1 = \int_0^{\frac{1}{\theta}} g(x)dx = \sum_{i=0}^{\infty} \int_{\frac{2^{i+1}}{\theta}}^{\frac{2^i}{\theta}} g(x)dx = 0$$

which is a contradiction.

3.4 A family $[p(D, \theta); \theta \in A]$ is called complete if, for any vector function $h(D)$ such that $\mathbb{E}_D[h; D] = 0$, $\forall \theta$, then $h = 0$. Show that if $[p(D; \theta) : \theta \in A]$ in complete. and there exists an MVU estimator, then this estimator is unique.

Suppose we have two MVU estimators, θ_1 and θ_2. Then $\mathbb{E}\left(\frac{\theta_1+\theta_2}{2}\right)$ is also unbiased. This estimator has variance

$$\mathbb{E}\left(\left(\frac{\theta_1+\theta_2}{2}\right)^T\left(\frac{\theta_1+\theta_2}{2}\right)\right) = \frac{var(\theta_1)}{4} + \frac{var(\theta_2)}{4} + \mathbb{E}\left(\frac{\theta_1^T\theta_2}{2}\right)$$

$var(\theta_1) = var(\theta_2)$ is a lower bound for this variance, which gives us an inequality. A little rearranging gives

$$\mathbb{E}(2\theta_1^T\theta_2) \geq \mathbb{E}(\theta_1^T\theta_1) + \mathbb{E}(\theta_2^T\theta_2)$$

We also know that

$$\mathbb{E}(2\theta_1^T\theta_2) \leq \mathbb{E}(\theta_1^T\theta_1) + \mathbb{E}(\theta_2^T\theta_2)$$

by examining the expectation of the positive quantity $(\theta_1 - \theta_2)^T(\theta_1 - \theta_2)$. Therefore both of these inequalities hold with equality, and $\mathbb{E}((\theta_1 - \theta_2)^T(\theta_1 - \theta_2)) = 0 \ \forall \theta$. By completeness, we have $(\theta_1 - \theta_2)^T(\theta_1 - \theta_2) = 0$ identically, which immediately gives $\theta_1 = \theta_2$, so the MVU estimator must be unique.

3.5 Let $\hat{\theta}_n$ be an unbiased estimator, so that $\mathbb{E}[\hat{\theta}_u] = \theta_0$. Define a biased one by $\hat{\theta}_b = (1+\alpha)\hat{\theta}_u$. Show that the range of α where the MSE of $\hat{\theta}_b$ is smaller than that of $\hat{\theta}_u$ is

$$\boldsymbol{-2 < -\frac{2\mathrm{MSE}(\hat{\theta_u})}{\mathrm{MSE}(\hat{\theta_u}) + \theta_0^2} < \alpha < 0}$$

We note that the MSE of $\hat{\theta}_u$ is

$$\mathbb{E}((\hat{\theta}_u - \theta_0)^2) = \mathbb{E}(\hat{\theta}_u^2) - \theta_0^2$$

by unbiasedness. Similarly, the MSE of $\hat{\theta}_b$ is

$$\mathbb{E}((\hat{\theta}_b - \theta_0)^2) = (1+\alpha)^2\mathbb{E}(\hat{\theta}_u^2) - (1+2\alpha)\theta_0^2 = (1+\alpha)^2\mathrm{MSE}(\hat{\theta_u}) + \alpha^2\theta_0^2$$

Therefore the condition we want occurs for exactly

$$(1+\alpha)^2\mathrm{MSE}(\hat{\theta_u}) + \alpha^2\theta_0^2 < \mathrm{MSE}(\hat{\theta_u})$$

which rearranges to

$$\alpha((\mathrm{MSE}(\hat{\theta_u}) + \theta_0^2)\alpha + 2\mathrm{MSE}(\hat{\theta_u})) < 0$$

This occurs only iff exactly one of α and $(\mathrm{MSE}(\hat{\theta_u}) + \theta_0^2)\alpha + 2 \cdot \mathrm{MSE}(\hat{\theta_u})$ is positive. Note also that $\alpha > 0$ directly implies $(\mathrm{MSE}(\hat{\theta_u}) + \theta_0^2)\alpha + 2 \cdot \mathrm{MSE}(\hat{\theta_u}) > 0$. Thus our proposed condition holds iff

$$-\frac{2\mathrm{MSE}(\hat{\theta_u})}{\mathrm{MSE}(\hat{\theta_u}) + \theta_0^2} < \alpha < 0$$

as required. The final leftmost inequality stems from the fact that

$$\frac{\mathrm{MSE}(\hat{\theta_u})}{\mathrm{MSE}(\hat{\theta_u}) + \theta_0^2} < 1$$

3.6 Show that for the setting of Problem 3.4, the optimal value of α is equal to

$$\boldsymbol{\alpha_* = -\frac{1}{1 + \frac{\theta_0^2}{var(\hat{\theta}_u)}}}$$

We note that our MSE for the biased estimator is a quadratic in α. Therefore we can just pick the unique value of α for which the derivative is 0, and we obtain the minimum possible MSE. The derivative is

$$2(\mathrm{MSE}(\hat{\theta_u}) + \theta_0^2)\alpha + 2\mathrm{MSE}(\hat{\theta_u})$$

which rearranges to the required result.

3.7 Show that the regularity condition for the Cramer-Rao lower bound holds true if the order of differentiation and integration can be interchanged.

Note that $\frac{\partial \log(p(x|\theta))}{\partial\theta} = \frac{1}{p(x|\theta)} \cdot \frac{\partial\, p(x|\theta)}{\partial\theta}$. This gives

$$\mathbb{E}\left(\frac{\partial \log(p(x|\theta))}{\partial\theta}\right) = \int_{x\in X\cap\theta} p(x|\theta)\cdot\frac{1}{p(x|\theta)}\cdot\frac{\partial\, p(x|\theta)}{\partial\theta}\, dx = \int_{x\in X\cap\theta} \frac{\partial\, p(x|\theta)}{\partial\theta}\, dx$$

where $X\cap\theta$ is the probability space created by intersecting the probability space X with the event θ, and μ is its associated probability measure. (We fix θ to be a single event - we are using the frequentist concept of a "true" value for θ)

Now, if we assume that differentiation and integration can be interchanged, then this is equal to

$$\frac{\partial}{\partial\theta}\int_{x\in X\cap\theta} p(x|\theta)\, d\mu$$

The expression being differentiated is just the integral of the probability of all events over the probability space, and so so is just equal to 1. Therefore the derivative is equal to 0, and thus the regularity condition holds, as required.

3.8 Derive the Cramer-Rao bound for the LS estimator, when the training data result from the linear model

$$\boldsymbol{y_n = \theta x_n + \eta_n, \quad n = 1, 2, ...,}$$

where $\boldsymbol{x_n}$ and $\boldsymbol{\eta_n}$ are i.i.d. samples of a zero mean random variable with variance $\boldsymbol{\sigma_x^2}$, and a Gaussian random variable with zero mean and variance $\boldsymbol{\sigma_\eta^2}$ respectively. Assume also that $\boldsymbol{x}$ and $\boldsymbol{\eta}$ are independent[sic]. Then show that the LS simulator achieves the CR bound only asymptotically.

Assume that the input variable X and output Y are observable and that the elements of $\mathbf{x}$ and $\boldsymbol{\eta}$ are all mutually independent. Therefore our pdf is separable as

$$f(\mathbf{a}, \mathbf{b}, \theta) = \prod_{n=1}^{N} p_x^n(a_n) p_\eta^n(b_n - \theta a_n)$$

where p_x^n is the marginal pdf of x_n and p_η^n is the marginal pdf of η_n. Taking the log gives

$$-\sum_{n=1}^{N} \frac{(b_n - \theta a_n)^2}{2\sigma_\eta^2} + \text{ terms independent of } \theta$$

Differentiating this twice w.r.t θ and multiplying through by -1 yields

$$\sum_{n=1}^{N} \frac{a_n^2}{\sigma_\eta^2}$$

The expectation of this quantity (as a multiple integral over all the a_n and b_n) is the Fisher information. We can make our lives easier by making the change of variables $c_n = b_n - \theta a_n$ and reframing our intgeral as being over the a_n and c_n instead. This separates the base pdf completely into a_n and c_n terms. Since the quantity we are taking the expectation of only contains a_n terms, the Fisher information immediately simplifies to

$$\sum_{n=1}^{N} \int_{-\infty}^{\infty} p_x^n(a_n) \frac{a_n^2}{\sigma_\eta^2} da_n$$

Since the x_n are i.i.d with zero mean and variance σ_x^2 this is just equal to

$$N \cdot \frac{\sigma_x^2}{\sigma_\eta^2}$$

Therefore the Cramer-Rao lower bound is

$$\frac{\sigma_\eta^2}{N \cdot \sigma_x^2}$$

Given a set of observations $\mathbf{x}, \mathbf{y}$, the value of $\hat{\theta}$ that minimises least squares is

$$\hat{\theta} = \frac{\mathbf{x} \cdot \mathbf{y}}{\mathbf{x} \cdot \mathbf{x}} = \frac{\mathbf{x} \cdot (\theta \mathbf{x} + \boldsymbol{\eta})}{\mathbf{x} \cdot \mathbf{x}} = \theta + \frac{\mathbf{x} \cdot \boldsymbol{\eta}}{\mathbf{x} \cdot \mathbf{x}}$$

This quantity has expectation θ since we can separate out the zero-mean $\boldsymbol{\eta}$ from the second term. Therefore the variance is just

$$\mathbb{E}\left(\frac{(\mathbf{x} \cdot \boldsymbol{\eta})^2}{(\mathbf{x} \cdot \mathbf{x})^2}\right)$$

Considering that the elements of η are mutually independent and have expectation zero, we can ignore cross terms in the expansion of the numerator, which allows us to simplify the above expression to

$$\mathbb{E}\left(\frac{\sum_{n=1}^{N}(x_n \eta_n)^2}{(\mathbf{x} \cdot \mathbf{x})^2}\right) = \sigma_\eta^2 \mathbb{E}\left(\frac{\mathbf{x} \cdot \mathbf{x}}{(\mathbf{x} \cdot \mathbf{x})^2}\right) = \mathbb{E}\left(\frac{\sigma_\eta^2}{\mathbf{x} \cdot \mathbf{x}}\right)$$

Note that $\mathbf{x} \cdot \mathbf{x}$ is the sum of N independent copies of $\mathbf{x}$, and thus by the strong law of large numbers

$$\frac{\mathbf{x} \cdot \mathbf{x}}{N} \xrightarrow{a.s.} \sigma_x^2$$

where the convergence is almost sure convergence. By the continuous mapping theorem we immediately get

$$\frac{N}{\mathbf{x}\cdot\mathbf{x}} \xrightarrow{a.s.} \frac{1}{\sigma_x^2}$$

Under the assumption that this family of random variables is uniformly integrable (which I think is true but showing it is effort) the expectations exist and converge to the expectation of the limit. TODO - work out how to get from a.s. convergence to L1 comvergence

3.9 Let us consider the regression model

$$\boldsymbol{y_n = \theta^T x_n + \eta_n, \quad n = 1, 2, 3, ..., n}$$

where the noise samples $\boldsymbol{\eta = [\eta_1, ..., \eta_N]^T}$ come from a zero-mean Gaussian random vector with covariance matrix $\boldsymbol{\Sigma_\eta}$. If X is our input matrix and y is our output vector, show that

$$\boldsymbol{\hat{\theta} = (X^T \Sigma_\eta^{-1} X)^{-1} X^T \Sigma_\eta^{-1} y}$$

is a sufficient estimate.

Here we treat the x_n as known quantities, and model the η_n as our only source of uncertainty. Using the relation $\mathbf{y} = \boldsymbol{\theta}^T\mathbf{x} + \mathbf{y}$ we can rewrite $\hat{\boldsymbol{\theta}}$ as

$$\theta + (X^T\Sigma_\eta^{-1}X)^{-1}X^T\Sigma_\eta^{-1}\eta$$

The covariance of this expression is just

$$(X^T\Sigma_\eta^{-1}X)^{-1}$$

The likelihood function here is given by

$$p(y = a) = p(\eta = a - \theta^T X) = \frac{1}{((2\pi)^d|\Sigma_\eta|)^{\frac{1}{2}}} \prod_{n=1}^{N} e^{\frac{-(a-\theta^T X)^T \Sigma_\eta^{-1}(a-\theta^T X)}{2}}$$

From this, we obtain the Fisher information matrix as simply $(X^T\Sigma_\eta^{-1}X)$, and parameter efficiency via Cramer-Rao follows trivially.

3.10 Assume a set of i.i.d. $\boldsymbol{X = \{x_1, x_2, ..., x_N\}}$ samples of a random variable[sic] with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma^2}$. Define also the quantities

$$\boldsymbol{S_\mu := \frac{1}{N}\sum_{n=1}^{N} x_n}$$

$$\boldsymbol{S_{\sigma^2} := \frac{1}{N}\sum_{n=1}^{N}(x_n - S_\mu)^2}$$

$$\bar{S}_{\sigma^2} := \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

Show that if μ is considered to be known, a sufficient statistic for σ^2 is $\bar{S}_{\sigma^2}$. Moreover in the case where both (μ, σ^2) are unknown, then a sufficient statistic is the pair (S_μ, S_{σ^2})

This is not well-posed in its current form. It does however make sense if we constrain the x_i to be drawn from a Gaussian distribution.

We note the the p.d.f of X is

$$p(X = x) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{\frac{-\sigma^2(x-\mu)^T(x-\mu)}{2}}$$

This can be written as

$$\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{\frac{-\sigma^2 N \bar{S}_{\sigma^2}}{2}}$$

which is a function of σ and $\bar{S}_{\sigma^2}$ only. Therefore by Fisher-Neyman we have the first result. For the second part, we note that the p.d.f may alternatively be written as

$$\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{\frac{-N\sigma^2(S_{\sigma^2} - 2\mu S_\mu + \mu^2 + S_\mu^2)}{2}}$$

which is solely a function of μ, σ^2, S_μ, and S_{σ^2}. Therefore we have the second result by Fisher-Neyman.

3.11 Show that solving the task

$$\textbf{minimise } L(\theta, \lambda) = \sum_{n=1}^{N}\left(y_n - \theta_0 - \sum_{i=1}^{l}\theta_i x_{ni}\right)^2 + \lambda\sum_{i=1}^{l}|\theta_i|^2$$

is equivalent to solving the task

$$\textbf{minimise } L(\theta, \lambda) = \sum_{n=1}^{N}\left((y_n - \bar{y}) - \sum_{i=1}^{l}\theta_i(x_{ni} - \bar{x}_i)\right)^2 + \lambda\sum_{i=1}^{l}|\theta_i|^2$$

and the estimate of θ_0 is given by

$$\hat{\theta}_0 = \bar{y} - \sum_{i=1}^{l}\hat{\theta}_i\bar{x}_i$$

The derivative of the first expression is necessarily 0 when evalutated at any minimiser $\hat{\theta}$. By evaluating the derivative with respect to θ_0 at the minimiser, we immediately get

$$\hat{\theta}_0 = \bar{y} - \sum_{i=1}^{l}\hat{\theta}_i\bar{x}_i$$

Substituting this into the expression to be minimised reveals that minimising over the other N components is exactly equivalent to solving the first constraint of the second optimisation problem. We already know that the minimiser to our first problem meets the second constraint of the second problem, and therefore will solve the problem in its entirety.

The second optimisation problem has a unique solution (can be seen by explicitly expressing as a quadratic form), so the above implication shows that the first problem also has a unique solution. Thus the two problems are exactly equivalent.

3.12 Consider the linear regression task

$$\boldsymbol{y_n = \theta_o + \eta_n}$$

where the $\boldsymbol{\eta_n}$ are i.i.d. zero-mean Gaussian variables of variance $\boldsymbol{\sigma_\eta^2}$. Show that the ridge regression estimator with parameter $\boldsymbol{\lambda}$, $\boldsymbol{\hat{\theta}_b(\lambda)}$, and the minimum variance unbiased estimator, $\boldsymbol{\hat{\theta}_{\mathrm{MVU}}}$, of $\boldsymbol{\theta_o}$ satisfy the following:

$\mathbf{MSE}(\boldsymbol{\hat{\theta}_b(\lambda)}) < \mathbf{MSE}(\boldsymbol{\hat{\theta}_{\mathrm{MVU}}})$ provided that

$$\begin{cases} \boldsymbol{\lambda \in (0, \infty)} & \boldsymbol{\theta_o^2 \leq \frac{\sigma_\eta^2}{N}} \\ \boldsymbol{\lambda \in \left(0, \frac{2\sigma_\eta^2}{\theta_o^2 - \frac{\sigma_\eta^2}{N}}\right)} & \boldsymbol{\theta_o^2 > \frac{\sigma_\eta^2}{N}} \end{cases}$$

Moreover, show that the minimum MSE performance for the ridge regression estimate is attained at $\boldsymbol{\lambda_* = \frac{\sigma_\eta^2}{\theta_o^2}}$.

We know that in this case with i.i.d. Gaussian noise and constant input that the MVU estimator is just $\bar{y}$ the mean of the sample outputs. We thus calculate

$$\mathrm{MSE}(\hat{\theta}_{\mathrm{MVU}}) = \mathbb{E}((\bar{y} - \theta_o)^2) = \mathbb{E}((\sum_{n=1}^{N} \frac{\eta_n}{N})^2)$$

Since the η_n are independent and have zero mean, the cross terms disappear and we are left with

$$\mathrm{MSE}(\hat{\theta}_{\mathrm{MVU}}) = (\sum_{n=1}^{N} \mathbb{E}(\frac{\eta_n^2}{N^2})) = N \cdot \frac{\sigma_\eta^2}{N^2} = \frac{\sigma_\eta^2}{N}$$

Similarly, we know from studying the ridge regression problem that $\hat{\theta}_b(\lambda)$ satisfies

$$(\lambda + N)\hat{\theta}_b(\lambda) = \sum_{n=1}^{N} y_n = N\theta_o + \sum_{n=1}^{N} \eta_n$$

and therefore

$$\hat{\theta}_b(\lambda) = \frac{N}{N+\lambda}\theta_o + \sum_{n=1}^{N} \frac{\eta_n}{N+\lambda}$$

We now have

$$\mathrm{MSE}(\hat{\theta}_b(\lambda)) = \mathbb{E}\left(\left(\frac{N}{N+\lambda}\theta_o + \sum_{n=1}^{N} \frac{\eta_n}{N+\lambda} - \theta_o\right)^2\right) = \mathbb{E}\left(\left(\sum_{n=1}^{N} \frac{\eta_n}{N+\lambda} - \frac{\lambda\theta_o}{N+\lambda}\right)^2\right)$$

We again use the fact that the η_n are independent and have zero mean to obtain

$$\mathrm{MSE}(\hat{\theta}_b(\lambda)) = \frac{\lambda^2\theta_o^2 + N\sigma_\eta^2}{(N+\lambda)^2}$$

Therefore the statement $\mathrm{MSE}(\hat{\theta}_b(\lambda)) < \mathrm{MSE}(\hat{\theta}_{\mathrm{MVU}})$ is equivalent to

$$\frac{\lambda^2\theta_o^2 + N\sigma_\eta^2}{(N+\lambda)^2} < \frac{\sigma_\eta^2}{N}$$

Multiplying through by each denominator reveals this to be equivalent to stating

$$N\lambda^2\theta_o^2 + N^2\sigma_\eta^2 < (N+\lambda)^2\sigma_\eta^2$$

or rather

$$(\sigma_\eta^2 - N\theta_o^2)\lambda^2 + 2\sigma_\eta^2 N\lambda > 0$$

Our outline of ridge regression already assumes that λ is positive. We can see by considering the sign of the quadratic term that if $\theta_o^2 \leq \frac{\sigma_\eta^2}{N}$ this expression is positive for all $\lambda \in (0, \infty)$. Similarly, if $\theta_o^2 > \frac{\sigma_\eta^2}{N}$, the expression is positive exactly for

$$\lambda \in \left(0, \frac{2\sigma_\eta^2}{\theta_o^2 - \frac{\sigma_\eta^2}{N}}\right)$$

as required. To calculate where the ridge regression estimate achieves minimum MSE, we take the derivative (with the help of the quotient rule) w.r.t. λ:

$$\frac{d(\mathrm{MSE}(\hat{\theta}_b(\lambda)))}{d\theta} = \frac{2\lambda\theta_o^2(N+\lambda)^2 - 2(\lambda^2\theta_o^2 + N\sigma_\eta^2)(N+\lambda)}{(N+\lambda)^4}$$

This becomes 0 in the case where

$$\lambda\theta_o^2(N+\lambda) = (\lambda^2\theta_o^2 + N\sigma_\eta^2)$$

which directly gives the optimal value of λ as

$$\lambda_* = \frac{\sigma_\eta^2}{\theta_o^2}$$

3.13 Consider the regression model given in problem 3.9, but with $\boldsymbol{\Sigma} = \boldsymbol{I}_N$. Compute the MSE of the predictions $\mathbb{E}[(\boldsymbol{y} - \hat{\boldsymbol{y}})^2]$, where $\boldsymbol{y}$ is the true response and $\hat{\boldsymbol{y}}$ is the predicted value, given a test point $\boldsymbol{x}$ and using the LS estimator

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

The LS estimator has been obtained via a set of N measurements, collected in the input matrix X and y. The expectation $\mathbb{E}[\cdot]$ is taken with respect to y, the training data $\mathcal{D}$, and the test points x.

There are multiple sources of uncertainty here, so we should lay out the problem carefully. In our test set, we have N (x, y)-pairs. We use these to calculate