Show, by differentiating the above loss, that the analytical solution is $wRidge$=(XTX+λI)−1XTy

$$L(y,\hat{y}) = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda||w||_2{}^2 = \sum_{i=1}^{N}(\hat{y}_i - y_i)^2 + \lambda||w||_2{}^2 = \sum_{i=1}^{N}(x_i^T w - y_i)^2 + \lambda||w||_2{}^2$$
$$= (Xw - y)^T(Xw - y) + \lambda w^T w$$

$$\frac{dL}{dw} = 2X^T(Xw - y) + \lambda(2w) = 0$$

$$X^T Xw - X^T y + \lambda w = 0$$

$$(X^T X + \lambda I)w = X^T y$$

$$w_{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

Bonus: Noise as a regularizer:

$X' = X * G$ where G is an uncorrelated noise with variance σ and mean 1.

I.E we scale each entry of X by a small amount of Gaussian noise: $x_{i,j} \rightarrow sx_{i,j}$ where $s \sim N(1,\sigma)$.

We get a different line for each choice of ϵ, so for OLS we need to find the expectation vector w which minimize the error: $\widehat{w} \sim argmin_w E_G[|y - (G * X)w|^2]$ (E_G marginalizes out the contributions of the noise).

Start with the expression inside the expectation:

$$|y - (G * X)w|^2 = (y - (G * X)w)^T(y - (G * X)w) = y^T y - 2y^T(G * X)w + w^T(G * X)^T(G * X)w$$

Let us define: $(G * X)^T(G * X) = M$

A single entry in M is: $m_{i,j} = \sum_k s_{ki} \cdot s_{kj} \cdot x_{ki} \cdot x_{kj}$ which in expectation is $E[M_{i,j}] = \sum_k E[s_{ki} \cdot s_{kj}] x_{ki} \cdot x_{kj}$.

If $i \neq j$ then s_{ki} and s_{kj} are independent and drawn from $N(1,\sigma)$, so $\sum_k E[s_{ki} \cdot s_{kj}] = 1$.

If $i = j$ then s_{ki} and s_{kj} are not independent and $E[s_{ki} \cdot \epsilon_{kj}] = E[s_{ki}^2]$, and by using abbreviated multiplication formula:

$$E[s_{ki}^2] = E[(s_{ki} - 1)^2 + 2s_{ki} - 1] = \sigma^2 + 2 - 1 = \sigma^2 + 1$$

So if 1 is a square matrix with a 1 in every entry:

$$E[M] = \left(1 + diag(\sigma^2)\right) * X^T X = X^T X + diag(\sigma^2) X^T X$$

And:

$$E[|y - (G * X)w|^2] = E[y^T y - 2y^T(G * X)w + w^T(G * X)^T(G * X)w] = y^T y - 2y^T(E[G] * X)w + w^T E[M]w$$
$$= y^T y - 2y^T Xw + w^T X^T Xw + w^T diag(\sigma^2) X^T Xw$$

And again by using abbreviated multiplication formula:

$$= |y - Xw|^2 + w^T diag(\sigma^2) X^T Xw = |y - Xw|^2 + \sigma^2 |\sqrt{diag(X^T X)} \cdot w^2|$$

So going back we need to find w such as: $\widehat{w} \sim argmin_w(|y - Xw|^2 + \sigma^2 \left|\sqrt{diag(X^TX)} \cdot w\right|^2)$ which is linear regression with some regulation term.

Because in ridge regularization we standardize the predictors it ensures that $\frac{1}{N} diag(X^TX) = I$. So for our expression:

$$|y - Xw|^2 + \sigma^2 \left|\sqrt{diag(X^TX)} \cdot w\right|^2 = |y - Xw|^2 + \sigma^2 \left|\sqrt{NI} \cdot w\right|^2 = |y - Xw|^2 + N\sigma^2 |w|^2$$

And we get: $\widehat{w} \sim argmin_w(|y - Xw|^2 + N\sigma^2|w|^2)$ which is ridge with $\lambda = N\sigma^2$.