Regularization and Condition Number

Part A

Ridge regression has a residual sum of squares (RSS) of the form:

$$\begin{aligned} RSS(\beta, X, y) &= ||y - X\beta||^2 + \gamma||\beta||^2 \\ &= \sum_i (y_i - (X\beta)_i)^2 + \gamma \sum_i \beta_i^2 \end{aligned} \tag{1}$$

The ordinary least squares (OLS) solution is found by minimizing this with respect to β. To do this, we note a few derivatives:

$$\frac{\partial \beta_i}{\partial \beta_j} = \delta_{ij} \tag{2}$$

$$\frac{\partial (X\beta)_i}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} \sum_k X_{ik}\beta_k = \sum_k X_{ik}\delta_{jk} = X_{ij} \tag{3}$$

Using this we find:

$$\begin{aligned} \frac{\partial RSS}{\partial \beta_j} = \sum_i \frac{\partial}{\partial \beta_j}(y_i - (X\beta)_i)^2 + \gamma \sum_i \frac{\partial \beta_i^2}{\partial \beta_j} &= 0 \\ \sum_i 2(y_i - (X\beta)_i)(-X_{ij}) + \gamma \sum_i 2\beta_i \delta_{ij} &= 0 \\ \sum_i (-X_{ji}^T y_i) + \sum_i X_{ji}^T (X\beta)_i + \gamma\beta_j &= 0 \\ -(X^T y)_j + (X^T X\beta)_j + \gamma\beta_j &= 0 \end{aligned} \tag{4}$$

Or if we convert back to matrix notation, and noting that $\gamma\beta = \gamma\mathbb{I}\beta$:

$$X^T X\beta - X^T y + \gamma\mathbb{I}\beta = 0$$

$$(X^T x + \gamma\mathbb{I})\beta = x^T y \tag{5}$$

$$\beta = (X^T X + \gamma\mathbb{I})^{-1} X^T y \tag{6}$$

This matches the solution we derived in class.

Part B

Let:

$$A = X^T X \tag{7}$$

Which is non-singular, real and positive definite.

Consider the matrix equation that matches onto the solution in part A:

$$a = A^{-1} b \tag{8}$$

Such that a = β, b = $X^T y$, and we consider $\gamma = 0$ for now. Doing this, we can relate errors in b (our data) to errors in a (the OLS parameter solution) such that:

$$a + \Delta a = (A + \Delta A)^{-1}(b + \Delta b) \tag{9}$$

Doing this, we note that if $\Delta A = 0$, then we can relate the uncertainties in a and b directly:

$$\Delta a = A^{-1} \Delta b \tag{10}$$

We can define the condition number for A as the upper bound on how inaccurate an approximate solution for a will be. That is, it tells us how magnified uncertainties in our data (underlying noise due to the fundamental probability distributions, for example) become as they propagate through to the solution. This can be defined as (for $\Delta A = 0$):

$$cond(A) = \underset{b,\Delta b}{max} \frac{||\Delta a||/||a||}{||\Delta b||/||b||} \tag{11}$$

To move on, we need to define the **matrix operator norm** for a matrix. This is defined as:

$$||A||_p = \underset{x}{max} \frac{||Ax||_p}{||x||_p} \tag{12}$$

Where $||x||_p$ is understood to be the p-norm for a vector, ($[\sum_i x_i^p]^{1/p}$) and $||A||_p$ the corresponding matrix p-norm. Thus, this is effectively a measure of how much A can stretch a vector. Moving on, we drop the p and assume that context identifies whether we are considering a matrix or vector p-norm.

Rewriting the condition number for A, noting that $\Delta a = A^{-1}\Delta b$ and $b = Aa$, we find:

$$\begin{aligned} cond(A) &= \underset{a,\Delta b}{max} \frac{||A^{-1}\Delta b||/||a||}{||\Delta b||/||Aa||} \\ &= \underset{a,\Delta b}{max} \frac{||A^{-1}\Delta b||}{||\Delta b||} \cdot \frac{||Aa||}{||a||} \\ &= \underset{\Delta b}{max} \frac{||A^{-1}\Delta b||}{||\Delta b||} \cdot \underset{a}{max} \frac{||Aa||}{||a||} \\ &= ||A^{-1}|| \cdot ||A|| \end{aligned} \tag{13}$$

Next we wish to show that this is greater than 1. To do this, we return to the definition of the matrix operator norm. Consider the norm applied to 2 matrices:

$$\begin{aligned} ||AB|| &= \max_x \frac{||ABx||}{||x||} \\ &= \max_x \frac{||A(Bx)||}{||x||}\frac{||Bx||}{||Bx||} \end{aligned} \tag{14}$$

If we let Bx = y, then:

$$||AB|| = \max_x \frac{||Ay||}{||y||}\frac{||Bx||}{||x||} \tag{15}$$

Because, by definition, $||A|| \geq ||Ax||/||x|| \; \forall x$, we can thus write this as:

$$||AB|| \leq ||A|| \cdot ||B|| \tag{16}$$

Applying this to the condition number, we find:

$$cond(A) = ||A^{-1}|| \cdot ||A|| \geq ||A^{-1}A|| = ||\mathbb{I}|| = 1 \tag{17}$$

Thus small condition numbers imply that the precision in our solution is only limited by the underlying precision in our data. For large conditions numbers, the problem is ill-conditioned and small uncertainties can propagate through to large errors in the final parameters.

It's also worth nothing that for the p-2 norm, we can show that a unitary matrix has unit norm ($||U|| = max||Ux||/||x|| = 1$) as they preserve standard L2 lengths. We can further show that:

$$\begin{aligned} ||UA|| &\leq ||U||||A|| = ||A|| \\ ||A|| &= ||U^TUA|| \leq ||U^T||||UA|| = ||UA|| \\ ||UA|| &= ||A|| \end{aligned} \tag{18}$$

And likewise for $||AU|| = ||A||$.

This means that the singular value decomposition of A (USV^T) can be written as:

$$||A|| = ||USV^T|| = ||SV^T|| = ||S|| \tag{19}$$

For a diagonal matrix (such as S = $diag(s_1, s_2, ..., s_n)$), we have:

$$||Sx||_2^2 = \sum_i |s_i|^2|x_i|^2 \leq (\max_i \; |s_i|^2)||x||^2 \tag{20}$$

And so it's easy to see that $||S|| = |s_{max}|$. For a real, positive definite matrix, (such as $A = X^TX$), the singular values correspond to the eigenvalues, λ. Furthermore:

$$\begin{aligned} A^{-1} &= (USV^T)^{-1} = (V^T)^{-1}S^{-1}U^{-1} \\ &= VS^{-1}U^T \end{aligned} \tag{21}$$

And so the eigenvectors of A^{-1} are simply the inverse of A. Thus, we have, for the p-2 norm:

$$cond(A) = ||A^{-1}|| \cdot ||A|| = \frac{1}{|\lambda_{min}|} \cdot |\lambda_{max}| \tag{22}$$

Part C

Now let's continue with the assumption that there is no error in A. We also revert back to the parameter space, and consider the errors in β instead of a. Consider the condition number:

$$cond(A) = \underset{b,\Delta b}{max} \frac{||\Delta\beta||/||\beta||}{||\Delta b||/||b||} \tag{23}$$

This is the maximum possible magnification of the underlying errors. Thus for arbitrary b and Δb, we have:

$$\frac{||\Delta\beta||/||\beta||}{||\Delta b||/||b||} \leq cond(A) \tag{24}$$

If we further assume $||\Delta b|| < ||b||$ we can arrive at:

$$\frac{||\Delta\beta||}{||\beta||} \leq \frac{||b||}{||\Delta b||} cond(A) \leq cond(A) \tag{25}$$

This means that our relative error is bounded only by the condition number of A. A well-conditioned matrix will always lead to converging solutions for the parameters, regardless of the method used to determine them.

Let's also consider the condition number if we vary A and hold b fixed. In this case we have $\beta + \Delta\beta = (A + \Delta A)^{-1} b$, or, solving for $\Delta\beta = -A^{-1}\Delta A(\beta + \Delta\beta)$. Thus we have:

$$\begin{aligned}
\frac{||\Delta\beta||/||\beta||}{||\Delta A||/||A||} &= \frac{||A||}{||\Delta A||} \frac{||A^{-1}\Delta A(\beta + \Delta\beta)||}{||\beta||} \\
&= \frac{||A||}{||\Delta A||} \frac{||A^{-1}\Delta A(\beta + \Delta\beta)||}{||\beta + \Delta\beta||} \frac{||\beta + \Delta\beta||}{||\beta||} \\
&\leq \frac{||A||}{||\Delta A||} ||A^{-1}\Delta A|| \frac{||\beta + \Delta\beta||}{||\beta||} \\
&\leq \frac{||A||}{||\Delta A||} ||A^{-1}||\ ||\Delta A|| \frac{||\beta + \Delta\beta||}{||\beta||} \\
&\leq ||A|| \cdot ||A^{-1}|| \frac{||\beta + \Delta\beta||}{||\beta||} \\
&\leq ||A|| \cdot ||A^{-1}|| \frac{||\beta||}{||\beta||} = ||A|| \cdot ||A^{-1}|| = cond(A)
\end{aligned} \tag{26}$$

This cannot be reduced further, and thus we find the same condition number as before. If we work from the second last line, we can thus show:

$$\begin{aligned} \frac{||\Delta\beta||/||\beta||}{||\Delta A||/||A||} &\leq cond(A)\frac{||\beta+\Delta\beta||}{||\beta||} \\ \frac{||\Delta\beta||}{||\beta+\Delta\beta||} &\leq \frac{||\Delta A||}{||A||}cond(A) \end{aligned} \tag{27}$$

Part D

Now we consider the full ridge regression solution for the inverse matrix. That is, we use:

$$B = A + \gamma\mathbb{I} \tag{28}$$

Such that $\beta = B^{-1}b$. Now consider the eigen-decomposition of $A = Q\Lambda Q^{-1}$, where $\Lambda = diag(\lambda_{max}, ..., \lambda_{min})$:

$$\begin{aligned} B &= Q\Lambda Q^{-1} + \gamma\mathbb{I}QQ^{-1} \\ &= Q\Lambda Q^{-1} + Q\gamma\mathbb{I}Q^{-1} \\ &= Q(\Lambda + \gamma\mathbb{I})Q^{-1} \end{aligned} \tag{29}$$

And so we see that the eigenvalues of B are simply those of A, plus γ. Also note that, because A is positive definite, its eigenvalues are all positive, and so this preserves the ordering of the eigenvalues: the largest (smallest) eigenvalue of B is $\lambda_{max(min)} + \gamma$. Thus:

$$\begin{aligned} cond(B) &= \frac{\lambda_{max}+\gamma}{\lambda_{min}+\gamma} \\ &= \left(\frac{\lambda_{max}}{\lambda_{min}}\right)\frac{1+\gamma/\lambda_{max}}{1+\gamma/\lambda_{min}} \\ &= cond(A)\frac{1+\gamma/\lambda_{max}}{1+\gamma/\lambda_{min}} \end{aligned} \tag{30}$$

Because $\lambda_{max} > \lambda_{min}$ we know that $\gamma/\lambda_{max} < \gamma/\lambda_{min}$ and so:

$$cond(B) \leq cond(A) \tag{31}$$