

Exercises

§9.7: Regularization and Feature Selection

Python: There is an optional starter script that supports Python-based problem(s) on this sheet. Use it as scaffolding: read the comments, complete the lines marked **TODO**, and compare the output with the mathematics requested in the problem(s).

Problem 1 Consider $g(w) = (w - 4)^2 + \lambda w^2$. (a) Find the ridge minimizer as a function of λ . (b) Compute it for $\lambda = 0, 1, 3$. (c) Explain why it does not become exactly zero for finite λ .

Problem 2 Consider $g(w) = (w - 4)^2 + \lambda|w|$. Derive the minimizer by considering the cases $w > 0$, $w < 0$, $w = 0$. Compute the solution for $\lambda = 2, 6, 8, 10$. At what value of λ does the coefficient first reach zero?

Problem 3 For $X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$, $\mathbf{y} = (1, 2, 3)^T$, and $\lambda = 1/3$, (a) compute the ordinary least-squares solution, (b)

compute the ridge solution from $(X^T X + P\lambda I)\mathbf{w} = X^T \mathbf{y}$, (c) compare the two coefficient norms, (d) compare the training SSE values.

Problem 4 Suppose a model uses $\hat{y} = wx$ and a ridge penalty λw^2 . Now replace the feature by $z = 100x$. (a) What coefficient $\tilde{w}$ produces identical predictions? (b) Compare $\tilde{w}^2$ with w^2 . (c) Explain why this demonstrates that ridge regularization depends on feature scale.

Problem 5 Suppose $X^T X$ has eigenvalues 100, 1, 0.001. Let $P\lambda = 1$. What are the eigenvalues of $X^T X + P\lambda I$? Compute the condition number before and after regularization. Explain the stabilization.

Problem 6 A model has $N = 20$ candidate predictors. (a) How many possible predictor subsets are there? (b) Why does exhaustive subset search become difficult as N grows? (c) Explain how lasso approaches feature selection differently. (d) Why should a lasso-selected feature not automatically be interpreted as causal?

Problem 7 Use Python and np to compute ridge coefficients for

```
x = np.array([[1., 0.],  
              [0., 1.],  
              [1., 1.]])
```

```
y = np.array([1., 2., 3.])
```

for $\lambda = 0, 1/3, 1$. Report the coefficient vector and its squared L_2 norm for each value. Describe the coefficient path.

Problem 8 Two standardized predictors are highly correlated. Explain qualitatively how ridge and lasso may behave differently. Which method is more likely to retain both predictors? Which is more likely to set one coefficient to zero? Why can the particular lasso-selected predictor be unstable across samples?