Foundations of Deep Learning **Fall 2024**

Exercise 10: Regularization

Lecturer: Aurelien Lucchi

Problem 1 (Regularization warm-up with Ridge Regression):

Consider a data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ that contains n datapoints $\mathbf{x}_i \in \mathbb{R}^d$ for $i = 1, \dots, n$, as well as the corresponding targets $y_i \in \mathbb{R}$. The ridge regression solution for a linear model is given by

$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in \mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2, \tag{1}$$

where $\lambda > 0$ and $\mathbf{y} = (y_1, \dots y_n) \in \mathbb{R}^n$.

With the singular value decomposition $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, show that the prediction $\hat{y} := \mathbf{X}\hat{\mathbf{w}}$ has the following form:

$$\hat{\mathbf{y}} = \sum_{i=1}^{d} \mathbf{u}_i \frac{d_i^2}{d_i^2 + \lambda} \mathbf{u}_i^\top \mathbf{y}. \tag{2}$$

How does the above differ from the predictions made by a non-regularized linear estimator? What is the effect of λ?

Problem 2 (Connection between early stopping and L^2 regularization):

Early stopping is a regularization method, used when training a learner with an iterative method, in which the parameters from an earlier iteration (rather than the last one) are returned. In most cases, the criterion used to decide the "stop time" is the error on the validation set.

Formally, we are interested in optimizing the risk function $\mathcal{R}(\mathbf{w})$. We assume we can approximate it using the second-order approximation around an optimal $\mathbf{w}^*$, i.e.

$$\mathcal{R}(\mathbf{w}) \approx \mathcal{R}(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H}(\mathbf{w} - \mathbf{w}^*). \tag{3}$$

Our goal will be to identify a connection between early stopping and L^2 regularization for a linear model with parameters $\mathbf{w}$. To do so, we will follow these steps:

a) Starting from the quadratic approximation of $\mathcal{R}(\mathbf{w})$, compute its gradient and write down the update rule for $\mathbf{w}_k$ given by gradient descent. Derive a recursive equation for the difference vector $\mathbf{w}_k - \mathbf{w}^*$.

b) Re-write the equation you derived in step 1 using the eigen-decomposition $\mathbf{H} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$.

c) Collapse the recursive equation (i.e., expand steps $k-1, \dots, 1, 0$) to get a simplified expression for $\mathbf{w}_k$. You should get to an expression that depends on the term $(\mathbf{w}_0 - \mathbf{w}^*)$. Then simplify the obtained expression by assuming that $\mathbf{w}_0 = \mathbf{0}$.

d) Now consider the regularizer risk $\mathcal{R}(\mathbf{w}) + \lambda\|\mathbf{w}\|_2^2$. Show that the first-order optimality condition can be written as

$$\mathbf{Q}^\top \mathbf{w} = \left[\mathbf{I} - \lambda(\mathbf{\Lambda} + \lambda\mathbf{I})^{-1}\right]\mathbf{Q}^\top \mathbf{w}^*. \tag{4}$$

Hint: recall that if $\mathbf{A}, \mathbf{B}$ are two invertible matrices, then $(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

e) Conclude by matching the two formulas and by deriving a connection between the weight decay factor λ, the learning rate η, and the (early) stopping time τ. (You can use approximations, e.g., make some assumptions about the eigenvalues relative to these parameters to show how they depend on one another.)