Exercises
Introduction to Machine Learning
Spring 2023

Institute for Machine learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause, Prof. Dr. Fanny Yang

Homework 1 (Linear Regression, Optimization, Regularization)

For questions, please refer to Moodle.
Released on **8 March 2023**

General Instructions

- Submission of solutions is not mandatory but solving the exercises is highly recommended.
- The master solution will be released next week.

Exercise 1: Convex functions

In the following exercise we consider real-valued functions $f : \mathrm{dom}(f) \to \mathbb{R}$, with $\mathrm{dom}(f) \subseteq \mathbb{R}^d$.

A function $f : \mathrm{dom}(f) \to \mathbb{R}$ is *convex* if (i) $\mathrm{dom}(f)$ is a convex set and (ii) for all $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}(f)$ and all $\lambda \in [0, 1]$, we have

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(x) + (1-\lambda) f(\boldsymbol{y}) \tag{1}$$

Geometrically, the condition means that the line segment connecting the points $(\boldsymbol{x}, f(\boldsymbol{x})), (\boldsymbol{y}, f(\boldsymbol{y})) \in \mathbb{R}^{d+1}$ lies point-wise above the graph of f; see Figure 1 below.

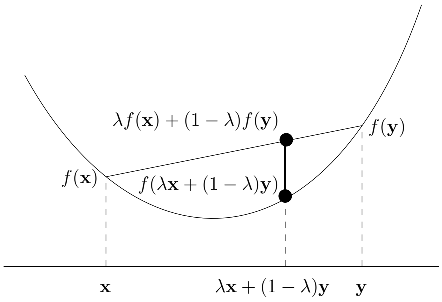
$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$
$f(\mathbf{y})$
$f(\mathbf{x})$
$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y})$
x
$\lambda\mathbf{x} + (1-\lambda)\mathbf{y}$
y

Figure 1: A convex function

Questions

(a) A norm is a function $f : \mathbb{R}^d \to \mathbb{R}$ which satisfies the following three properties:

- $f(x) > 0$ for all $x \neq 0$
- $f(\theta x) = |\theta| f(x)$ for all $\theta \in \mathbb{R}$ and $x \in \mathbb{R}^d$
- $f(x+y) \leq f(x) + f(y)$ for all $x, y \in \mathbb{R}^d$

Show that any valid norm f is a convex function using the definition of convexity in Equation 1.

Solution: Let f be a norm satisfying the three properties above. We have

$$f\left(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}\right) \leq f(\lambda \boldsymbol{x}) + f((1-\lambda)\boldsymbol{y}) = \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) \tag{2}$$

where for the first step we used property (iii) and in the second step we used property (ii).

Consider now the function $f(x,y) = x^2 + y^2$. The graph of f is the unit paraboloid in $\mathbb{R}^3$ which looks convex. However, verifying the condition in (1) directly is somewhat cumbersome.

To address this problem, we develop better ways to check convexity if the function under consideration is differentiable. In particular, suppose that $\text{dom}(f)$ is open and that f is differentiable, i.e. the gradient

$$\nabla f(\boldsymbol{x}) := \left(\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_p}(x)\right)$$

exists at every point $\boldsymbol{x} \in \text{dom}(f)$. Then, we will prove that an easier condition to verify convexity is the following:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) \tag{3}$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \text{dom}(f)$. Geometrically, this means that for all $\boldsymbol{x} \in \text{dom}(f)$, the graph of f lies above its tangent hyperplane at the point $(\boldsymbol{x}, f(\boldsymbol{x}))$; see Figure 2 below.

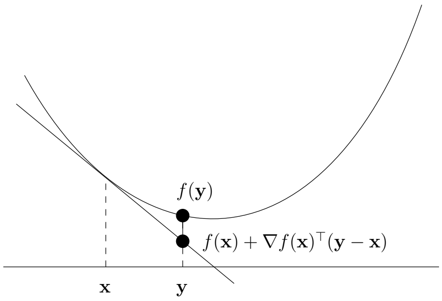
$f(\mathbf{y})$
$f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x})$
x
y

Figure 2: First-order characterization of convexity

Questions

(b) Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable. Prove that f is convex if and only if $\text{dom}(f)$ is convex and

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) \tag{4}$$

holds for all $\boldsymbol{x}, \boldsymbol{y} \in \text{dom}(f)$.

Solution: Suppose that f is convex, meaning that for $t \in [0,1]$ we have

$$f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) = f((1-t)\boldsymbol{x} + t\boldsymbol{y}) \leq (1-t)f(\boldsymbol{x}) + tf(\boldsymbol{y}) = f(\boldsymbol{x}) + t(f(\boldsymbol{y}) - f(\boldsymbol{x})) \tag{5}$$

Dividing by t and using differentiability at $\mathbf{x}$, we get

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}) + \frac{f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{t} \\ &= f(\mathbf{x}) + \frac{\nabla f(\mathbf{x})^\top t(\mathbf{y} - \mathbf{x}) + r(t(\mathbf{y} - \mathbf{x}))}{t} \\ &= f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{r(t(\mathbf{y} - \mathbf{x}))}{t} \end{aligned}$$

where the error term $r(t(\mathbf{y} - \mathbf{x}))/t$ goes to 0 as $t \to 0$. The inequality

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

follows.

Now suppose this inequality holds for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$, let $\lambda \in [0,1]$, and define $\mathbf{z} := \lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in \mathrm{dom}(f)$ (by convexity of $\mathrm{dom}(f)$). Then we have

$$\begin{aligned} f(\mathbf{x}) &\geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\top(\mathbf{x}-\mathbf{z}), \\ f(\mathbf{y}) &\geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\top(\mathbf{y}-\mathbf{z}). \end{aligned}$$

After multiplying the first inequality by λ and the second one by $(1-\lambda)$, the gradient terms cancel in the sum of the two inequalities, and we get

$$\lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}) \geq f(\mathbf{z}) = f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y})$$

(c) Suppose that f is differentiable and convex. Prove that $\boldsymbol{x}^\star$ is a global minimum of f if and only if $\nabla f(\boldsymbol{x}^\star) = 0$.

Solution: Suppose that $\nabla f(\boldsymbol{x}^\star) = 0$. According to Exercise 1b), we have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}^\star) + \nabla f(\boldsymbol{x}^\star)^\top(\boldsymbol{y}-\boldsymbol{x}^\star) = f(\boldsymbol{x}^\star) \tag{6}$$

for all $\boldsymbol{y} \in \mathrm{dom}(f)$, so $\boldsymbol{x}^\star$ is a global minimum.

The converse is also true and does not even require convexity. Suppose that $\nabla f(\boldsymbol{x}^\star)_i \neq 0$ for some i. For $t \in \mathbb{R}$, we define $\boldsymbol{x}(t) = \boldsymbol{x}^\star + t\mathbf{e}_i$, where $\mathbf{e}_i$ is the i-th unit vector. For $|t|$ sufficiently small, we have $\boldsymbol{x}(t) \in \mathrm{dom}(f)$ since $\mathrm{dom}(f)$ is open. Let $z(t) = f(\boldsymbol{x}(t))$. By the chain rule, $z'(0) = \nabla f(\boldsymbol{x}^\star)^\top\mathbf{e}_i = \nabla f(\boldsymbol{x}^\star)_i \neq 0$. Hence, z decreases in one direction as we move away from $t = 0$, and this yields $f(\boldsymbol{x}(t)) < f(\boldsymbol{x}^\star)$ for some t, so $\boldsymbol{x}^\star$ is not a global minimum.

(d) Show that $f(x,y) = x^2 + y^2$ is a convex function and that the point $(0,0)$ is a global minimum.

Solution: For $f(x_1, x_2) = x_1^2 + x_2^2$ we have that the gradient is equal to

$$\nabla f(x_1, x_2) = (2x_1, 2x_2) \tag{7}$$

hence substituting into Equation 4 we have

$$y_1^2 + y_2^2 \geq x_1^2 + x_2^2 + 2x_1(y_1 - x_1) + 2x_2(y_2 - x_2) \tag{8}$$

which after some rearranging of terms is equivalent to

$$(y_1 - x_1)^2 + (y_2 - x_2)^2 \geq 0 \tag{9}$$

hence according to Exercise 1b), f is convex.

Moreover, the gradient of f at the point $(0,0)$ is $\nabla f(0,0) = \mathbf{0}$ and f is differentiable on $\mathrm{dom}(f) = \mathbb{R}$, hence according to Exercise 1c), the point is a global minimum.

A third way to verify convexity is through:

$$\nabla^2 f(x) \succeq 0 \tag{10}$$

Geometrically, this means that the graph of f has non-negative curvature everywhere and hence looks like a bowl. We will come back to this when discussing ridge regression in Exercise 3.

Questions

(e) A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is called α-strongly convex for some $\alpha > 0$, if for any points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ one has

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{\alpha}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

If f is twice differentiable, an equivalent condition is that for any point $x \in \mathbb{R}^d$, one has

$$\nabla^2 f(x) \succeq \alpha I$$

which means $\nabla^2 f(\boldsymbol{x}) - \alpha I$ is positive semi-definite for all $\boldsymbol{x} \in \mathbb{R}^d$. Prove that a strongly convex function admits a unique minimizer in $\mathbb{R}^d$.
Hint: This is not an easy exercise. First prove that $f(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$ to show that there is some minimizer.

Solution: First, let us prove that the function f is coercive, i.e., $\lim_{\|\boldsymbol{x}\| \to \infty} f(\boldsymbol{x}) = \infty$. In the definition of strong convexity, by setting $\boldsymbol{x} = 0$ we get

$$f(\boldsymbol{y}) \geq f(0) + \nabla f(0)^\top \boldsymbol{y} + \frac{\alpha}{2}\|\boldsymbol{y}\|^2 \geq f(0) - \|\nabla f(0)\| \cdot \|\boldsymbol{y}\| + \frac{\alpha}{2}\|\boldsymbol{y}\|^2,$$

where we used the Cauchy-Schwartz inequality: $\nabla f(0)^\top \boldsymbol{y} \geq -\|\nabla f(0)\| \cdot \|\boldsymbol{y}\|$. The right-hand side of the equation above is a quadratic function of $\|\boldsymbol{y}\|$ with a positive coefficient for the second degree term. Thus, it goes to infinity as $\|\boldsymbol{y}\| \to \infty$. Hence, f also goes to infinity.

Next, we prove that f has a global minimum. Denote by $s = \inf_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}) < \infty$. Then, by the definition of the infimum, there exists a sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ such that $f(\boldsymbol{x}_n) \to s$. We claim that this sequence is bounded: otherwise, there would be a subsequence such that $\|\boldsymbol{x}_{n_i}\| \to \infty$. But as f is coercive, $f(\boldsymbol{x}_{n_i}) \to \infty$, contradicting $f(\boldsymbol{x}_{n_i}) \to s < \infty$. Hence, the sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ is inside some bounded set. By compactness, we obtain that there exists a convergent subsequence. As f is continuous, the f value of this subsequence converges as well, meaning that the infimum is attained. That is, $\exists \boldsymbol{x}_\infty : f(\boldsymbol{x}_\infty) = s = \inf f(\boldsymbol{x})$.

Finally, we prove uniqueness. If $\boldsymbol{x}$ and $\boldsymbol{y}$ were two distinct global minima for f, then, by strong convexity, we have

$$f\left(\frac{\boldsymbol{x} + \boldsymbol{y}}{2}\right) < \frac{1}{2}(f(\boldsymbol{x}) + f(\boldsymbol{y})) = \min f,$$

which results in a contradiction.

Exercise 2: Linear regression

In the lecture we have learned how to fit an affine function to data by performing linear regression. In the tutorial on Friday we will discuss how to fit more general nonlinear functions to data. The goal of this exercise is to solidify our understanding of some of the concepts that have been touched upon.

Consider a dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R} \times \mathbb{R}$ and the hypothesis space of affine functions $H = \{w_0 + w_1 x : w_0, w_1 \in \mathbb{R}\}$. The error of a solution $f \in H$, i.e., $f(x) = w_0 + w_1 x$ for some $w_0, w_1 \in \mathbb{R}$ is given by

$$L(w_0, w_1) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i)^2. \tag{11}$$

Questions

(a) Let us for a moment consider the simpler case where we fix $w_0 = 0$. Compute the optimal linear fit to the data by computing $w_1^* = \arg\min_{w_1 \in \mathbb{R}} L(0, w_1)$.

Solution: We start by computing the first derivative

$$\frac{\partial L(0, w_1)}{\partial w_1} = \frac{1}{n} \sum_{i=1}^{n} (w_1 x_i - y_i) x_i$$

and the second derivative

$$\frac{\partial^2 L(0, w_1)}{\partial w_1^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2.$$

We observe that the second derivative is greater or equal to zero for all $w_1 \in \mathbb{R}$. Thus, $L(0, w_1)$ is convex and we can find its global minimum (if it exists) by setting its first derivative to zero

$$\frac{1}{n} \sum_{i=1}^{n} (w_1 x_i - y_i) x_i = 0$$

$$\Leftrightarrow w_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

If not all x_i are equal to zero, w_1^* is given by the equation above.

(b) Prove that, for $n \geq 2$ and $x_i \neq x_j$ for $i \neq j$, (11) is a strictly convex function with respect to $\boldsymbol{w} = (w_0, w_1)$.

Solution: We are going to show that (11) is strictly convex by showing that its Hessian $\nabla^2 L(w_0, w_1)$ is positive definite. First we compute the gradient $\nabla L(w_0, w_1)$

$$\nabla L(w_0, w_1) = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^{n} w_1 x_i + w_0 - y_i \\ \sum_{i=1}^{1} (w_1 x_i + w_0 - y_i) x_i \end{pmatrix}$$

Then, we can derive the gradient to obtain the Hessian

$$\nabla^2 L(w_0, w_1) = \frac{1}{n} \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} = \frac{1}{n} \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}$$

Now, a 2×2 symmetric matrix is positive definite if and only if both its trace and determinant are strictly greater than zero. First, we show that the trace is strictly positive:

$$\mathrm{Tr}(\nabla^2 L(w_0, w_1)) = \frac{1}{n}(n + \sum_{i=1}^{n} x_i^2) > 0 \;\; \forall x_i \in \mathbb{R} \tag{12}$$

Next, we show that the determinant is strictly positive:

$$n^2 \det(\nabla^2 L(w_0, w_1)) = n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2 > 0 \tag{13}$$

since $x_i - x_j \neq 0$ for $i \neq j$.

(c) The unique global minimum of a strictly convex function can be computed by setting its gradient to zero. Compute the gradient

$$\nabla L(w_0, w_1) = \begin{pmatrix} \frac{\partial L(w_0,w_1)}{\partial w_0} \\ \frac{\partial L(w_0,w_1)}{\partial w_1} \end{pmatrix}. \tag{14}$$

Solution: We get

$$\nabla L\left(w_0, w_1\right) = \left(\begin{array}{c} \frac{\partial L(w_0,w_1)}{\partial w_0} \\ \frac{\partial L(w_0,w_1)}{\partial w_1} \end{array} \right) = \left(\begin{array}{c} \frac{1}{n}\sum_{i=1}^{n} w_0 + w_1 x_i - y_i \\ \frac{1}{n}\sum_{i=1}^{n} \left(w_0 + w_1 x_i - y_i\right) x_i \end{array} \right).$$

(d) Compute the optimal parameters $(w_0^*, w_1^*) = \arg\min_{w_0, w_1 \in \mathbb{R}} L(w_0, w_1)$ by solving the linear system of equations obtained by setting (14) to zero, i.e., $\nabla L(w_0, w_1) = 0$.

Solution: From setting $\frac{\partial L(w_0,w_1)}{\partial w_0}$ to zero we get

$$w_0^* = \frac{\sum_{i=1}^{n} y_i - w_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - w_1 \bar{x}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Now, plugging w_0^* into $\frac{\partial L(w_0,w_1)}{\partial w_1}$ and setting the resulting expression to zero we get

$$w_1^* = \frac{\sum_{i=1}^{n} x_i y_i - x_i \bar{y}}{\sum_{i=1}^{n} x_i^2 - x_i \bar{x}}$$

which can be rewritten (by adding zeros to both enumerator and denominator) to

$$\begin{aligned} w_1^* &= \frac{\sum_{i=1}^{n}\left(x_i y_i - x_i \bar{y}\right) + \overbrace{\sum_{i=1}^{n}\left(\overline{xy} - \bar{x} y_i\right)}^{=0}}{\sum_{i=1}^{n}\left(x_i^2 - x_i \bar{x}\right) + \underbrace{\sum_{i=1}^{n}\left(\bar{x}^2 - \bar{x} x_i\right)}_{=0}} \\ &= \frac{\sum_{i=1}^{n}\left(\bar{x} - x_i\right)\left(\bar{y} - y_i\right)}{\sum_{i=1}^{n}\left(\bar{x} - x_i\right)^2} \\ &= \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}. \end{aligned}$$

Another way of writing (11) is in matrix notation

$$L(\boldsymbol{w}) = \frac{1}{2n}\|\boldsymbol{y} - \Phi\boldsymbol{w}\|^2, \tag{15}$$

where $\boldsymbol{y} = (y_1, \dots, y_n)^T$ is the vector of target values, $\boldsymbol{w} = (w_0, w_1)^T$ is our weight vector and

$$\Phi = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \tag{16}$$

is the data matrix.

For $n \geq 2$ different observations (15) is a strictly convex function and can be minimized by setting its gradient

$$\nabla L(\boldsymbol{w}) = \frac{1}{n}(\Phi^T\Phi\boldsymbol{w} - \Phi^T\boldsymbol{y}) \tag{17}$$

to zero.

The benefit of (15) is that it straightforwardly generalizes to multiple inputs $x_i \in \mathbb{R}^d$ using

$$\Phi = \begin{pmatrix} 1 & x_{11} & \dots & x_{1d} \\ 1 & x_{21} & \dots & x_{2d} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nd} \end{pmatrix}, \tag{18}$$

where we have one row per data point and one column per input.

Questions

(e) Provide necessary conditions for Φ such that $\Phi^T\Phi$ is invertible.

Solution: A necessary and sufficient condition for invertibility of $\Phi^\top\Phi$ is:

$$\Phi^\top\Phi \text{ is invertible} \iff \text{rank}(\Phi) = d + 1 \tag{19}$$

We first prove that $\text{rank}(\Phi) = d + 1 \implies \Phi^\top\Phi$ is invertible by defining a function f as

$$f(x) := x^\top\Phi^\top\Phi x = \|\Phi x\|^2$$

which is positive semidefinite. Moreover, f vanishes when $\Phi x = 0$. If Φ has full column rank, then $\Phi x = 0$ implies that $x = 0$, i.e., f is positive definite. Hence, $\Phi^T\Phi$ is positive definite and thus invertible.

We next show that $\Phi^\top\Phi$ is invertible $\implies \text{rank}(\Phi) = d + 1$ which concludes the proof. If Φ has not full column rank then there exists $v \in \mathbb{R}^d \backslash \{0\}$ with $\Phi v = 0$. It follows that $\Phi^\top\Phi v = 0$ and hence $\Phi^\top\Phi$ is not invertible.

(f) Show that if $\Phi^T\Phi$ is invertible, then $\boldsymbol{w}^* = (\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{y}$ is the unique minimum of $L(\boldsymbol{w})$ in (15).

Solution: First, we compute the gradient:

$$\nabla L(\boldsymbol{w}) = \frac{1}{n}(\Phi^T\Phi\boldsymbol{w} - \Phi^T\boldsymbol{y}) \tag{20}$$

and by setting it to zero we get the following solution:

$$w^* = (\Phi^T\Phi)^{-1}\Phi^T \boldsymbol{y} \tag{21}$$

Now, all is left to show is that the Hessian of $L(w)$ is positive definite, this implies strict convexity which in turn implies unique minimum.

First, we derive the Hessian:

$$\nabla_w^2 L(w) = \frac{1}{n}\Phi^\top\Phi \tag{22}$$

Now if Φ is full rank then $\Phi^\top\Phi$ is positive definite (see solution of exercise 2e) and hence the minimum is unique.

(g) Show that for $n < d+1$ the regression problem

$$\min_{w \in \boldsymbol{R}^{d+1}} \|\boldsymbol{y} - \Phi \boldsymbol{w}\|^2 \tag{23}$$

does not admit a unique solution.

Solution: If $n < d+1$, we have $\text{rank}(\Phi) = \min(n, d+1) = n$. Further, we know $\text{rank}\left(\Phi^T\Phi\right) = \text{rank}(\Phi)$. By the rank-nullity theorem we have

$$\dim\left(\ker\left(\Phi^T\Phi\right)\right) = d + 1 - \text{rank}\left(\Phi^T\Phi\right) = d + 1 - n > 0$$

We have seen that our objective $\|y - \Phi w\|^2$ is convex. Thus, each minimal weight vector $w^* \in \arg\min_w \|\boldsymbol{y} - \Phi w\|^2$ satisfies $\nabla \|\boldsymbol{y} - \Phi w^*\|^2 = 0 \iff \Phi^T\Phi w^* = \Phi^T y$. Now, given a minimal weight vector w^*, we have $\Phi^T\Phi\left(w^* + u\right) = \Phi^T y$ for all $u \in \ker\left(\Phi^T\Phi\right)$. As $\dim\left(\ker\left(\Phi^T\Phi\right)\right) \geq 1$, there are infinitely many such $u \neq 0$.

Next, recall the gradient descent update as discussed in lecture: $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta\nabla_{\boldsymbol{w}} L(\boldsymbol{w}^t)$. We would like to compare the computational complexity of the closed-form solution $\boldsymbol{w}^* = (\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{y}$ against the one of the gradient descent algorithm. For the next exercises you may use the contraction inequality as discussed during lecture,

$$\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|_2 \leq \overbrace{\|I - \eta\Phi^\top\Phi\|_{op}}^{:=\rho} \|\boldsymbol{w}^t - \boldsymbol{w}^*\|_2$$

Questions

(h) Prove that if $\Phi^\top\Phi$ is full rank and the learning rate satisfies $\eta < \frac{2}{\lambda_{max}(\Phi^\top\Phi)}$, where λ_{max} is the largest eigenvalue, then $\rho < 1$. Conclude that gradient descent converges to the optimal solution w^* as $t \to \infty$.

Hint: remember that an equivalent characterization of the operator norm in terms of the eigenvalues of A is given by $\|A\|_{op} = \max\{|\lambda| : \lambda \text{ eigenvalue of } A\}$

Solution:

First, we note that if $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of the matrix $\Phi^\top \Phi$, then the eigenvalues of $I - \eta\Phi^\top\Phi$ are exactly $1-\eta\lambda_1, \ldots, 1-\ \eta\lambda_d$. Now, using the characterisation of operator norm given in the hint, we have

$$\begin{aligned}\rho &= \max\left\{\left|\lambda_{\max}\left(\boldsymbol{I} - \eta\Phi^\top\Phi\right)\right|, \left|\lambda_{\min}\left(\boldsymbol{I} - \eta\Phi^\top\Phi\right)\right|\right\}\\ &= \max\left\{\left|1 - \eta\lambda_{\min}\left(\Phi^\top\Phi\right)\right|, \left|1 - \eta\lambda_{\max}\left(\Phi^\top\Phi\right)\right|\right\}.\end{aligned}$$

Under a smaller step size $\eta \le \frac{1}{\lambda_{\max}\left(\Phi^\top\Phi\right)}$, both of the terms in the expression are non-negative and hence we can drop the absolute value function

$$\begin{aligned}\rho &= \max\left\{1 - \eta\lambda_{\min}\left(\Phi^\top\Phi\right), 1 - \eta\lambda_{\max}\left(\Phi^\top\Phi\right)\right\}\\ &= 1 - \eta\lambda_{\min}\left(\Phi^\top\Phi\right)\end{aligned}$$

Therefore, $\rho < 1$, as we have assumed that $\Phi^\top\Phi$ is full rank and hence $\lambda_{\min}\left(\Phi^\top\Phi\right) > 0$. On the other hand, if $\eta > \frac{1}{\lambda_{\max}\left(\Phi^\top\Phi\right)}$, we have

$$\begin{aligned}\rho &= \max\left\{\left|1 - \eta\lambda_{\min}\left(\Phi^\top\Phi\right)\right|, \left|1 - \eta\lambda_{\max}\left(\Phi^\top\Phi\right)\right|\right\}\\ &= \max\left\{1 - \eta\lambda_{\min}\left(\Phi^\top\Phi\right), \eta\lambda_{\min}\left(\Phi^\top\Phi\right) - 1, \eta\lambda_{\max}\left(\Phi^\top\Phi\right) - 1\right\}\\ &= \max\left\{1 - \eta\lambda_{\min}\left(\Phi^\top\Phi\right), \eta\lambda_{\max}\left(\Phi^\top\Phi\right) - 1\right\}\end{aligned}$$

If we want $\rho < 1$, two conditions shall be satisfied: (i) $1 - \eta\lambda_{\min}\left(\Phi^\top\Phi\right) < 1$, which is already fulfilled, as we assumed $\lambda_{\min}\left(\Phi^\top\Phi\right) > 0$, and (ii) $\eta\lambda_{\max}\left(\Phi^\top\Phi\right) - 1 < 1$, which holds if $\eta < \frac{2}{\lambda_{\max}\left(\Phi^\top\Phi\right)}$. Hence, it is clear that both conditions are satisfied under our assumptions.

Finally, observe that if $\rho < 1$, the upper bound on $\left\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\right\|_2$ converges linearly to zero as $t \to \infty$.

(i) Assume that the stepsize η is such that $\|I - \eta\Phi^T\Phi\|_{op} < 1$. Compute the number of gradient steps τ and the overall complexity required to obtain a solution $\boldsymbol{w}^\tau$ that satisfies $\|\boldsymbol{w}^\tau - \boldsymbol{w}^*\| < \varepsilon$, where $\boldsymbol{w}^\tau$ is the parameter vector computed by gradient descent after τ steps.

Solution: Let $\rho = \left\|I - \eta\Phi^T\Phi\right\|_{op}$. We want to find τ such that $\|w^\tau - w^*\| < \epsilon$. Applying the contraction inequality to $\|w^\tau - w^*\|$ τ times gives $\|w^\tau - w^*\| \le \rho^\tau \left\|w^0 - w^*\right\|$. Because $\|w^\tau - w^*\| \le \rho^\tau \left\|w^0 - w^*\right\|$ we have that $\|w^\tau - w^*\| < \epsilon$ when $\rho^\tau \left\|w^0 - w^*\right\| < \epsilon$. Bringing $\left\|w^0 - w^*\right\|$ to the other side and taking the logarithm on both sides yields

$$\begin{aligned}&\rho^\tau \left\|w^0 - w^*\right\| < \epsilon\\ \Leftrightarrow &\rho^\tau < \frac{\epsilon}{\|w^0 - w^*\|}\\ \Leftrightarrow &\tau \log\rho < \log\epsilon - \log\left\|w^0 - w^*\right\|\\ \Leftrightarrow &\tau > \frac{\log\epsilon - \log\|w^0 - w^*\|}{\log\rho},\end{aligned}$$

where in the last line the inequality is flipped because $\log\rho$ is smaller than zero for $0 < \rho < 1$. Let's denote $p = d + 1$. Computing the gradient at a given step $t+1$, i.e., computing $\nabla L\left(w^t\right) = \frac{1}{n}\left(\Phi^T\Phi w^t - \Phi^T y\right), O\left(p^2\right)$ operations to compute $\Phi^T\Phi w^t$ and once $O\left(np^2 + np\right) = O\left(np^2\right)$ operations to compute $\Phi^T\Phi$ and $\Phi^T y$ at the beginning of the algorithm. Subtracting

$\eta\nabla L\left(w^t\right)$ to w^t has negligible cost of $O(p)$. Therefore, in order to compute a solution w^τ that satisfies $\|w^\tau - w^*\| < \epsilon$ gradient descent requires $\tau = \left\lceil \frac{\log\epsilon - \log\|w^0 - w^*\|}{\log\rho} \right\rceil$ steps, resulting in a computational complexity of $O\left(np^2 + \frac{\log\epsilon - \log\|w^0-w^*\|}{\log\rho}p^2\right)$.

(j) For the linear regression loss function defined in Equation 15, prove that $L(w^{t+1}) < L(w^t)$ for small enough stepsize η.

Solution: Without loss of generality, let us assume that $\nabla L\left(\boldsymbol{w}^t\right)$ is nonzero (otherwise the gradient descent makes no move, and we are already at a stationary point). Using the multivariate Taylor expansion, we have

$$\begin{aligned} L\left(\boldsymbol{w}^{t+1}\right) &= L\left(\boldsymbol{w}^t - \eta\nabla L\left(\boldsymbol{w}^t\right)\right) \\ &= L\left(\boldsymbol{w}^t\right) - \eta\left\langle \nabla L\left(\boldsymbol{w}^t\right), \nabla L\left(\boldsymbol{w}^t\right)\right\rangle + R\left(-\eta\nabla L\left(\boldsymbol{w}^t\right)\right) \\ &= L\left(\boldsymbol{w}^t\right) - \eta\left\|\nabla L\left(\boldsymbol{w}^t\right)\right\|^2 + R\left(-\eta\nabla L\left(\boldsymbol{w}^t\right)\right) \end{aligned} \tag{24}$$

where $R:\mathbb{R}^d \to \mathbb{R}$ is the remainder term. This term is of $o\left(\eta\left\|\nabla L\left(\boldsymbol{w}^t\right)\right\|\right)$, i.e.,

$$\lim_{\eta\to 0}\frac{R\left(-\eta\nabla L\left(\boldsymbol{w}^t\right)\right)}{\eta\left\|\nabla L\left(\boldsymbol{w}^t\right)\right\|} = 0.$$

Then, by definition of the limit, for all $\epsilon > 0$ there exists $\eta > 0$ small enough such that

$$\left|\frac{R\left(-\eta\nabla L\left(\boldsymbol{w}^t\right)\right)}{\eta\left\|\nabla L\left(\boldsymbol{w}^t\right)\right\|}\right| \le \epsilon$$

or

$$\left|R\left(-\eta\nabla L\left(\boldsymbol{w}^t\right)\right)\right| \le \epsilon\eta\left\|\nabla L\left(\boldsymbol{w}^t\right)\right\|.$$

Applying this to (24) and picking $\epsilon < \left\|\nabla L\left(\boldsymbol{w}^t\right)\right\|$ we get

$$\begin{aligned} L\left(\boldsymbol{w}^{t+1}\right) &= L\left(\boldsymbol{w}^t\right) - \eta\left\|\nabla L\left(\boldsymbol{w}^t\right)\right\|^2 + R\left(-\eta\nabla L\left(\boldsymbol{w}^t\right)\right) \\ &\le L\left(\boldsymbol{w}^t\right) - \eta\left\|\nabla L\left(\boldsymbol{w}^t\right)\right\|^2 + \epsilon\eta\left\|\nabla L\left(\boldsymbol{w}^t\right)\right\| \\ &= L\left(\boldsymbol{w}^t\right) + \eta\left\|\nabla L\left(\boldsymbol{w}^t\right)\right\|\left(\epsilon - \left\|\nabla L\left(\boldsymbol{w}^t\right)\right\|\right) \\ &< L\left(\boldsymbol{w}^t\right), \end{aligned}$$

(k) Now, say you are free to choose a constant stepsize. What is the minimum number of iterations τ required to obtain a solution $\boldsymbol{w}^\tau$ that satisfies $\|\boldsymbol{w}^\tau - \boldsymbol{w}^*\| < \varepsilon$? How does it depend on the maximum and minimum eigenvalues λ_{max}, λ_{min} of the matrix $\Phi^T\Phi$?

Solution: In order to achieve the minimum number of required iterations τ, we first need to determine the optimal learning rate η. The operator norm associated with the L2 norm is the spectral norm. Thus, we have

$$\rho = \left\|I - \eta\Phi^T\Phi\right\| = \max\left(\left|1-\eta\lambda_{\max}\right|, \left|1-\eta\lambda_{\min}\right|\right)$$

where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and the largest eigenvalue of $\Phi^T\Phi$. Now, in order to determine the optimal choice of η we first derive an expression for ρ as a function of η. We have

$$\begin{aligned}\rho(\eta) &= \max\left(\left|1-\eta\lambda_{\max}\right|, \left|1-\eta\lambda_{\min}\right|\right)\\ &= \max\left(1-\eta\lambda_{\max}, \eta\lambda_{\max}-1, 1-\eta\lambda_{\min}, \eta\lambda_{\min}-1\right)\\ &= \max\left(\eta\lambda_{\max}-1, 1-\eta\lambda_{\min}\right)\end{aligned}$$

where the last equality holds because for $0 \leq \lambda_{\min} \leq \lambda_{\max}$ we have $\eta\lambda_{\max}-1 > \eta\lambda_{\min}-1$ and $1-\eta\lambda_{\min} > 1-\eta\lambda_{\max}$. Therefore, we have either $\rho = \eta\lambda_{\max}-1$, i.e., $\eta\lambda_{\max}-1 \geq 1-\eta\lambda_{\min}$ which is equivalent to $\eta \geq \frac{2}{\lambda_{\min}+\lambda_{\max}}$, or $\rho = 1-\eta\lambda_{\min}$, i.e., $1-\eta\lambda_{\min} > \eta\lambda_{\max}-1$ which is equivalent to $\eta < \frac{2}{\lambda_{\min}+\lambda_{\max}}$. As a result we get

$$\rho(\eta) = \begin{cases} 1-\eta\lambda_{\min} & \text{for } 0 < \eta < \frac{2}{\lambda_{\min}+\lambda_{\max}},\\ \eta\lambda_{\max}-1 & \text{for } \frac{2}{\lambda_{\min}+\lambda_{\max}} \leq \eta < \frac{1}{\lambda_{\max}}\end{cases},$$

which is a piecewise affine function in η that attains its minimum at $\eta^* = \frac{2}{\lambda_{\min}+\lambda_{\max}}$. The corresponding value of ρ is

$$\begin{aligned}\rho\left(\eta^*\right) &= 1-\frac{2\lambda_{\min}}{\lambda_{\min}+\lambda_{\max}}\\ &= \frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\min}+\lambda_{\max}}.\end{aligned}$$

Plugging the optimal $\rho\left(\eta^*\right)$ into the expression for τ from exercise (h) yields

$$\tau = \left\lceil \frac{\log\epsilon - \log\left\|w^0-w^*\right\|}{\log\left(\lambda_{\max}-\lambda_{\min}\right)-\log\left(\lambda_{\min}+\lambda_{\max}\right)} \right\rceil$$

(l) Compare the computational complexity of gradient descent to the one required to solve the linear system of equations $\Phi^T\Phi\boldsymbol{w}^* = \Phi^T\boldsymbol{y}$ in closed form.

Solution: The complexity of gradient descent is $O\left(np^2+\tau p^2\right)$. In comparison, the complexity of computing w^* in closed form, which is dominated by the cost of the matrix inversion $\left(\Phi^T\Phi\right)^{-1}$ is in $O\left(p^3\right)$. Furthermore, w^* can be also computed by solving the linear system of equations $\Phi^T\Phi w = \Phi^T y$, which is dominated by the $O\left(np^2\right)$ cost of computing $\Phi^T\Phi$ as solving the resulting $p \times p$ linear system of equations is in $O\left(p^2\right)$ and the computation of $\Phi^T y$ in $O(np)$. As a consequence, in terms of computational complexity it is best to compute w^* by solving the linear system of equations $\Phi^T\Phi w = \Phi^T y$ that is obtained by setting the gradient to zero. Comparing the cost of gradient descent to the cost of computing the closed form by multiplying with the inverse $\left(\Phi^T\Phi\right)^{-1}$ shows that gradient descent can be preferable if p is significantly larger than $n+\tau$.

Exercise 3: Bias variance tradeoff

In the lecture we have seen the bias variance decomposition of the expected squared error. In this exercise, we will first derive this decomposition in the case of linear regression and we will then use it to study ridge regression. Consider a dataset with fixed data matrix $\Phi \in \mathbb{R}^{n\times(d+1)}$ as shown in Equation 18 and noisy labels $\boldsymbol{y} \in \mathbb{R}^n$ with $y_i = \boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+\epsilon_i$. We define the noise ϵ_i such that $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$. We assume that we have used this training data to train a linear regressor with estimator $\hat{\boldsymbol{w}}$. We now consider the expected squared error of this estimate defined as $\mathbb{E}_\epsilon\left[\|\boldsymbol{w}-\hat{\boldsymbol{w}}\|^2\right]$. In this exercise, $\|\cdot\|$ always refers to the Euclidean norm. Remember the definitions of bias and variance as given in the lecture:

$$\mathrm{Bias}(\hat{\boldsymbol{w}})^2 = \|\boldsymbol{w} - \mathbb{E}_\epsilon[\hat{\boldsymbol{w}}]\|^2 \tag{25}$$

$$\mathrm{Var}(\hat{\boldsymbol{w}}) = \mathbb{E}_\epsilon\left[\|\hat{\boldsymbol{w}} - \mathbb{E}_\epsilon[\hat{\boldsymbol{w}}]\|^2\right] \tag{26}$$

Questions

(a) Prove that $\mathbb{E}_\epsilon\left[\|\boldsymbol{w} - \hat{\boldsymbol{w}}\|^2\right] = \mathrm{Bias}(\hat{\boldsymbol{w}})^2 + \mathrm{Var}(\hat{\boldsymbol{w}})$.

Solution:

$$\begin{aligned}\mathbb{E}_\epsilon\left[\|\boldsymbol{w} - \hat{\boldsymbol{w}}\|^2\right] &= \mathbb{E}_\epsilon\left[\|\boldsymbol{w} - \mathbb{E}_\epsilon[\hat{\boldsymbol{w}}] + \mathbb{E}_\epsilon[\hat{\boldsymbol{w}}] - \hat{\boldsymbol{w}}\|^2\right] \\ &= \mathbb{E}_\epsilon\left[\|\boldsymbol{w} - \mathbb{E}_\epsilon[\hat{\boldsymbol{w}}]\|^2\right] + 2(\boldsymbol{w} - \mathbb{E}_\epsilon[\hat{\boldsymbol{w}}])^\top \underbrace{\mathbb{E}_\epsilon\left[(\mathbb{E}_\epsilon[\hat{\boldsymbol{w}}] - \hat{\boldsymbol{w}})\right]}_{=0} + \mathbb{E}_\epsilon\left[\|\mathbb{E}_\epsilon[\hat{\boldsymbol{w}}] - \hat{\boldsymbol{w}}\|^2\right] \\ &= \|\boldsymbol{w} - \mathbb{E}_\epsilon[\hat{\boldsymbol{w}}]\|^2 + \mathbb{E}_\epsilon\left[\|\hat{\boldsymbol{w}} - \mathbb{E}_\epsilon[\hat{\boldsymbol{w}}]\|^2\right] = \mathrm{Bias}(\hat{\boldsymbol{w}})^2 + \mathrm{Var}(\hat{\boldsymbol{w}})\end{aligned}$$

We can see that even if our estimator $\hat{\boldsymbol{w}}$ is unbiased, we can still incur a large error if it has high variance. Similarly, even if the estimator has small variance, we can incur a large error if it is strongly biased. This is commonly referred to as the bias variance tradeoff.

Remember that while bias drives underfitting, variance is related to overfitting. During training, we aim to find the model with the lowest expected generalization error which requires joint optimization over these two components. As seen in the lecture, regularization is frequently used to control model complexity which reduces variance at the expense of increased bias. Ridge regression is a regularized version of linear regression. The optimization problem with parameter $\lambda > 0$ is given by Equation 27 which is optimizing over the loss we have seen in Exercise 2 but with an added regularization term $\lambda\|\boldsymbol{w}\|^2$.

$$\boldsymbol{w}^*_{\text{ridge}} = \arg\min_{\boldsymbol{w}} L_{\text{ridge}}(\boldsymbol{w}) = \arg\min_{\boldsymbol{w}} \left[\sum_{i=1}^{n}(y_i - \boldsymbol{w}^\top \phi(\boldsymbol{x}_i))^2 + \lambda\|\boldsymbol{w}\|^2\right] \tag{27}$$

Questions

(b) Prove that the solution to Equation 27 is unique for any matrix Φ by showing that L_{ridge} is strictly convex.

Solution: In matrix notation, we can write $L_{\text{ridge}}(\boldsymbol{w}) = \|\Phi\boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{w}\|^2$ (analogously to what we did for unregularized linear regression). We first compute the gradient $\nabla_{\boldsymbol{w}} L_{\text{ridge}}(\boldsymbol{w})$ by using the chain rule as:

$$\nabla_{\boldsymbol{w}} L_{\text{ridge}}(\boldsymbol{w}) = 2(\Phi\boldsymbol{w} - \boldsymbol{y})^\top\Phi + 2\lambda\boldsymbol{w}^\top = 2\boldsymbol{w}^\top\left(\Phi^\top\Phi + \lambda I_d\right) - 2\boldsymbol{y}^\top\Phi$$

Moreover, we can obtain the Hessian $\nabla^2_{\boldsymbol{w}} L_{\text{ridge}}(\boldsymbol{w})$ by differentiating $\nabla_{\boldsymbol{w}} L_{\text{ridge}}(\boldsymbol{w})$ again:

$$\nabla^2_{\boldsymbol{w}} L_{\text{ridge}}(\boldsymbol{w}) = 2\left(\Phi^\top\Phi + \lambda I_d\right).$$

Now, to prove that L_{ridge} is strictly convex we need that for some $\alpha > 0$:

$$\nabla^2_w L_{\text{ridge}}(w) - \alpha I_d \succeq 0 \tag{28}$$

and for $\alpha = 2\lambda$ the above condition is equivalent to:

$$2\Phi^\top\Phi \succeq 0 \tag{29}$$

which holds for any Φ since for any non-zero vector $\boldsymbol{v} \in \mathbb{R}^d$, we have $\boldsymbol{v}^\top\Phi^\top\Phi\boldsymbol{v} = \|\Phi v\|^2 \geq 0$.

Hence, L_{ridge} is 2λ-strongly convex, and from Exercise 1e) we can conclude that L_{ridge} has a unique minimizer.

In the following, we want to develop a better understanding of the bias variance tradeoff by comparing ridge regression to unregularized linear regression.

Questions

(c) Derive the closed-form solution $\boldsymbol{w}^*_{\text{ridge}}$ by computing the unique minimizer defined in Equation 27.

Solution: We know from task (b) that $\nabla_{\boldsymbol{w}} L_{\text{ridge}}(\boldsymbol{w}) = 2\boldsymbol{w}^\top\left(\Phi^\top\Phi + \lambda I_d\right) - 2\boldsymbol{y}^\top\Phi$ and by setting this to 0, we find:

$$\boldsymbol{w}^*_{\text{ridge}} = \left(\Phi^\top\Phi + \lambda I_d\right)^{-1}\Phi^\top\boldsymbol{y}$$

Note that the regularization term in the optimization problem admits a solution for any matrix Φ even in the case where $\Phi^\top\Phi$ is not be invertible.

(d) Show that its bias is given by $\text{Bias}(\boldsymbol{w}^*_{\text{ridge}}) = \|\lambda\left(\Phi^\top\Phi + \lambda I_d\right)^{-1}\boldsymbol{w}\|$. How does it compare to the bias of unregularized linear regression?

Solution: Remember that the bias depends on the noisy labels $\boldsymbol{y}$ with noise ϵ as defined in the beginning of Exercise 3. We find

$$\begin{aligned}
\text{Bias}(\boldsymbol{w}^*_{\text{ridge}}) &= \|\boldsymbol{w} - \mathbb{E}_\epsilon[\boldsymbol{w}^*_{\text{ridge}}]\| \\
&= \|\boldsymbol{w} - \mathbb{E}_\epsilon\left[\left(\Phi^\top\Phi + \lambda I_d\right)^{-1}\Phi^\top\boldsymbol{y}\right]\| \\
&= \|\boldsymbol{w} - \left(\Phi^\top\Phi + \lambda I_d\right)^{-1}\Phi^\top\mathbb{E}_\epsilon[\boldsymbol{y}]\| \\
&= \|\boldsymbol{w} - \left(\Phi^\top\Phi + \lambda I_d\right)^{-1}\Phi^\top\Phi\boldsymbol{w}\| \\
&= \|\left(\Phi^\top\Phi + \lambda I_d\right)^{-1}\left(\Phi^\top\Phi + \lambda I_d\right)\boldsymbol{w} - \left(\Phi^\top\Phi + \lambda I_d\right)^{-1}\Phi^\top\Phi\boldsymbol{w}\| \\
&= \|\lambda\left(\Phi^\top\Phi + \lambda I_d\right)^{-1}\boldsymbol{w}\|
\end{aligned}$$

(e) We next focus on the variance of the unregularized estimator as derived in Exercise 2. Remember that this estimator is defined as $\boldsymbol{w}^* = (\Phi^\top\Phi)^{-1}\Phi^\top\boldsymbol{y}$ assuming that Φ is full rank. Show that its variance is given by

$$\text{Var}(\boldsymbol{w}^*) = \sum_{i=1}^{m}\frac{\sigma^2}{\sigma_i^2}$$

where the σ_i are the singular values of Φ and $m = \min(n, d+1)$.

Hint: Remember that $\text{Var}(\hat{\boldsymbol{w}}) = \mathbb{E}_\epsilon\left[\|\hat{\boldsymbol{w}} - \mathbb{E}_\epsilon[\hat{\boldsymbol{w}}]\|^2\right]$ *and use the fact that* $A = Tr(A)$ *if* A *is a real number and that the trace is invariant under cyclic permutations* $Tr(ABC) = Tr(BCA) = Tr(CAB)$.

Solution: We begin by showing that $\text{Var}(\boldsymbol{w}^*) = \mathbb{E}_\epsilon\left[\|(\Phi^\top\Phi)^{-1}\Phi^\top\epsilon\|^2\right]$.

$$\begin{aligned}
\text{Var}(\boldsymbol{w}^*) &= \mathbb{E}_\epsilon\left[\|\boldsymbol{w}^* - \mathbb{E}_\epsilon[\boldsymbol{w}^*]\|^2\right] \\
&= \mathbb{E}_\epsilon\left[\|(\Phi^\top\Phi)^{-1}\Phi^\top\boldsymbol{y} - \mathbb{E}_\epsilon[(\Phi^\top\Phi)^{-1}\Phi^\top\boldsymbol{y}]\|^2\right] \\
&= \mathbb{E}_\epsilon\left[\|(\Phi^\top\Phi)^{-1}\Phi^\top\boldsymbol{y} - (\Phi^\top\Phi)^{-1}\Phi^\top\mathbb{E}_\epsilon[\boldsymbol{y}]\|^2\right] \\
&= \mathbb{E}_\epsilon\left[\|(\Phi^\top\Phi)^{-1}\Phi^\top(\Phi\boldsymbol{w}+\epsilon) - (\Phi^\top\Phi)^{-1}\Phi^\top\mathbb{E}_\epsilon[(\Phi\boldsymbol{w}+\epsilon)]\|^2\right] \\
&= \mathbb{E}_\epsilon\left[\|(\Phi^\top\Phi)^{-1}\Phi^\top(\Phi\boldsymbol{w}+\epsilon) - (\Phi^\top\Phi)^{-1}\Phi^\top\Phi\boldsymbol{w}\|^2\right] \\
&= \mathbb{E}_\epsilon\left[\|(\Phi^\top\Phi)^{-1}\Phi^\top\epsilon\|^2\right]
\end{aligned}$$

If we expand this term and define the empirical covariance matrix $\hat{\Sigma} = \Phi^T\Phi$, we get

$$\begin{aligned}
\mathbb{E}_\epsilon\left[\|(\Phi^\top\Phi)^{-1}\Phi^\top\epsilon\|^2\right] &= \mathbb{E}_\epsilon\left[\epsilon^\top\Phi\left((\Phi^\top\Phi)^{-1}\right)^\top(\Phi^\top\Phi)^{-1}\Phi^\top\epsilon\right] \\
&= \mathbb{E}_\epsilon\left[\epsilon^\top\Phi(\Phi^\top\Phi)^{-1}(\Phi^\top\Phi)^{-1}\Phi^\top\epsilon\right] \\
&= \mathbb{E}_\epsilon\left[\epsilon^\top\Phi\hat{\Sigma}^{-1}\hat{\Sigma}^{-1}\Phi^\top\epsilon\right]
\end{aligned}$$

As this is a real number, we can replace it with its trace (see hint). Using the invariance of the trace under cyclic permutations, we find

$$\epsilon^\top\Phi\hat{\Sigma}^{-1}\hat{\Sigma}^{-1}\Phi^\top\epsilon = \text{Tr}(\epsilon^\top\Phi\hat{\Sigma}^{-1}\hat{\Sigma}^{-1}\Phi^\top\epsilon) = \text{Tr}(\epsilon\epsilon^\top\Phi\hat{\Sigma}^{-1}\hat{\Sigma}^{-1}\Phi^\top)$$

Taking the expectation and exploiting the invariance of the trace under cyclic permutations, we finally arrive at

$$\begin{aligned}
\text{Var}(\boldsymbol{w}^*) &= \mathbb{E}_\epsilon\left[\text{Tr}(\epsilon\epsilon^\top\Phi\hat{\Sigma}^{-1}\hat{\Sigma}^{-1}\Phi^\top)\right] \\
&= \text{Tr}\left(\mathbb{E}_\epsilon\left[\epsilon\epsilon^\top\right]\Phi\hat{\Sigma}^{-1}\hat{\Sigma}^{-1}\Phi^\top\right) \\
&= \sigma^2\text{Tr}\left(\Phi\hat{\Sigma}^{-1}\hat{\Sigma}^{-1}\Phi^\top\right) \\
&= \sigma^2\text{Tr}\left(\Phi^\top\Phi\hat{\Sigma}^{-1}\hat{\Sigma}^{-1}\right) \\
&= \sigma^2\text{Tr}\left(\hat{\Sigma}\hat{\Sigma}^{-1}\hat{\Sigma}^{-1}\right) \\
&= \sigma^2\text{Tr}\left(\hat{\Sigma}^{-1}\right) \\
&= \sum_{i=1}^{d+1}\frac{\sigma^2}{\sigma_i^2}
\end{aligned}$$

where we have used that $\text{Var}(\epsilon_i) = \sigma^2$ and σ_i are the singular values of Φ. The last equality follows from the fact that $\text{Tr}(\hat{\Sigma}^{-1})$ is equal to the sum of its eigenvalues and the eigenvalues of $\hat{\Sigma}^{-1}$ are equal to the inverse of the eigenvalues of $\hat{\Sigma}$, which in turn are equal to the singular values of Φ squared.

(f) Similarly, it can be shown that the ridge estimator has variance

$$\text{Var}(\boldsymbol{w}^*_{\text{ridge}}) = \sigma^2\sum_{i=1}^{m}\frac{\sigma_i^2}{(\sigma_i^2+\lambda)^2}$$

where the σ_i are the singular values of Φ and $m = \min(n, d+1)$. Compare the variance of the unregularized estimator to that of the ridge estimator.

Solution: We can see that for $\lambda = 0$ we have $\mathrm{Var}(\boldsymbol{w}^*) = \mathrm{Var}(\boldsymbol{w}^*_{\text{ridge}})$. For $\lambda > 0$, we find that $\mathrm{Var}(\boldsymbol{w}^*_{\text{ridge}}) < \mathrm{Var}(\boldsymbol{w}^*)$ and in the extreme case where $\lambda \to \infty$, $\mathrm{Var}(\boldsymbol{w}^*_{\text{ridge}}) \to 0$.

(g) How does the choice of λ affect bias and variance? How do bias and variance behave as $\lambda \to 0$ and $\lambda \to \infty$?

Solution: To summarize, we have the following:

$$\begin{aligned}
\mathrm{Bias}(\boldsymbol{w}^*) &= 0 \\
\mathrm{Var}(\boldsymbol{w}^*) &= \sum_{i=1}^{m} \frac{\sigma^2}{\sigma_i^2} \\
\mathrm{Bias}(\boldsymbol{w}^*_{\text{ridge}}) &= \|\lambda \left(\Phi^\top \Phi + \lambda I_d\right)^{-1} \boldsymbol{w}\| \\
\mathrm{Var}(\boldsymbol{w}^*_{\text{ridge}}) &= \sigma^2 \sum_{i=1}^{m} \frac{\sigma_i^2}{(\sigma_i^2 + \lambda)^2}
\end{aligned}$$

As discussed in the lecture, regularization lowers the variance at the expense of introducing a bias. To see this, we will examing the limits $\lambda \to 0$ and $\lambda \to \infty$.

As $\lambda \to 0$, we find that $\boldsymbol{w}^*_{\text{ridge}} \to \boldsymbol{w}^*$ and hence $\mathrm{Bias}(\boldsymbol{w}^*_{\text{ridge}}) \to 0$ and $\mathrm{Var}(\boldsymbol{w}^*_{\text{ridge}}) \to \mathrm{Var}(\boldsymbol{w}^*)$.

For $\lambda \to \infty$, we can see by computing the limits that $\mathrm{Bias}(\boldsymbol{w}^*_{\text{ridge}}) \to \|\boldsymbol{w}\|$ and $\mathrm{Var}(\boldsymbol{w}^*_{\text{ridge}}) \to 0$.

Hence, increasing λ corresponds to reducing the variance and increasing the bias simultaneously while decreasing λ reduces the bias but increases the variance. This tradeoff corresponds to the tradeoff between underfitting (low variance, high bias) and overfitting (low bias, high variance) and hence the choice of λ has to be considered carefully when fine-tuning a model. We have to jointly minimize over the bias squared and the variance of our estimator to find the the solution with the lowest squared error.

Exercise 4: Gradient Descent for Linear Regression (Bonus)

In this exercise, we are going to prove that under mild conditions the gradient descent algorithm for ordinary linear regression problems converges to the solution with minimum norm. This is a very good exercise to practice your linear algebra skills. To help you prove this argument, we have divided the complete proof into smaller chunks. As in the lecture, suppose $X \in \mathbb{R}^{n \times d}$ is the data matrix and $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector. The goal is to find a vector $\boldsymbol{w} \in \mathbb{R}^d$ such that $L(\boldsymbol{w}) := \frac{1}{2}\|X\boldsymbol{w} - \boldsymbol{y}\|^2$ is minimized. For this, we use the gradient descent algorithm: starting from an initial vector $\boldsymbol{w}^0$, the iterates of the gradient descent algorithm for a step size η are

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k - \eta \nabla L(\boldsymbol{w}^k), \quad k = 0, 1, \dots$$

(a) By computing the gradient of L, confirm that

$$\boldsymbol{w}^{k+1} = (I - \eta X^\top X)\boldsymbol{w}^k + \eta X^\top \boldsymbol{y}.$$

Solution: We have seen in the lecture that

$$\nabla L(\boldsymbol{w}^k) = X^\top X \boldsymbol{w}^k - X^\top y$$

and if we insert this into the equation for the gradient descent algorithm, we find that:

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k - \eta(X^\top X\boldsymbol{w}^k - X^\top y) = (I - \eta X^\top X)\boldsymbol{w}^k + \eta X^\top \boldsymbol{y}$$

(b) By using induction on k, prove that

$$\boldsymbol{w}^k = \underbrace{(I - \eta X^\top X)^k \boldsymbol{w}^0}_{\text{(A)}} + \underbrace{\eta \left(\sum_{j=0}^{k-1} (I - \eta X^\top X)^j \right) X^\top \boldsymbol{y}}_{\text{(B)}} \tag{30}$$

Solution: Remember that an induction proof consists of two steps: we first show that a given equation holds for the base case (here $k = 0$) and we then show that if it holds for any given k, it must also hold for the next step at $k + 1$.

For $k = 0$, we find that part (B) in Equation (30) is 0 due to the empty set of summation indices j and we hence find that:

$$\boldsymbol{w}^0 = (I - \eta X^\top X)^0 \boldsymbol{w}^0 = \boldsymbol{w}^0$$

Assuming that Equation (30) holds for k, we can prove it for $k + 1$ by plugging it into the gradient descent equation and verifying that Equation (30) holds for $k + 1$ as well:

$$\begin{aligned}
\boldsymbol{w}^{k+1} &= (I - \eta X^\top X)\boldsymbol{w}^k + \eta X^\top \boldsymbol{y} \\
&= (I - \eta X^\top X) \left((I - \eta X^\top X)^k \boldsymbol{w}^0 + \eta \left(\sum_{j=0}^{k-1} (I - \eta X^\top X)^j \right) X^\top \boldsymbol{y} \right) + \eta X^\top \boldsymbol{y} \\
&= (I - \eta X^\top X)^{k+1} \boldsymbol{w}^0 + \eta \left(\sum_{j=0}^{k-1} (I - \eta X^\top X)^{j+1} \right) X^\top \boldsymbol{y} + \eta X^\top \boldsymbol{y} \\
&= (I - \eta X^\top X)^{k+1} \boldsymbol{w}^0 + \eta \left(\sum_{j=1}^{k} (I - \eta X^\top X)^{j} \right) X^\top \boldsymbol{y} + \eta X^\top \boldsymbol{y} \\
&= (I - \eta X^\top X)^{k+1} \boldsymbol{w}^0 + \eta \left(\sum_{j=0}^{k} (I - \eta X^\top X)^{j} \right) X^\top \boldsymbol{y}
\end{aligned}$$

From (b) it is clear that powers of the matrix $I - \eta X^\top X$ play an important role in understanding what happens to $\boldsymbol{w}^k$ when k is large. It is usual to look at the eigenvalues of a matrix when studying its powers. Hence, we start by the SVD of $X = U\Sigma V^\top$, where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{d\times d}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{n\times d}$ is a rectangular diagonal matrix with non-negative real numbers $\sigma_1, \ldots, \sigma_n$ on its diagonal. *From this part onwards, we focus on the over-parameterized case where $n < d$.*

(c) Verify that the eigenvalue decomposition of $I - \eta X^\top X$ is $V(I - \eta\Lambda)V^\top$, where $\Lambda \in \mathbb{R}^{d\times d}$ is a diagonal matrix whose first n diagonal entries are σ_i^2 and the rest are zero.

Solution: As $X^\top X = V\Sigma^\top \Sigma V^\top$ and $I = VV^\top$ (since V is orthogonal), we can write

$$I - \eta X^\top X = V(I - \eta \Sigma^\top \Sigma)V^\top = V(I - \eta\Lambda)V^\top,$$

where Λ is the matrix as described in the problem. Note that since V is orthogonal and $I - \eta\Lambda$ is diagonal, it follows that we have just derived the eigendecomposition of $I - \eta X^\top X$.

(d) Denote by $\sigma_{\max} := \max \sigma_i$. Observe that if $\eta \leq 1/\sigma_{\max}^2$, all eigenvalues of $I - \eta X^\top X$ will be non-negative.

Solution: The entries on the diagonal (and hence, the eigenvalues) of $I - \eta\Lambda$ are $1 - \eta\sigma_i^2$. Hence, if $\eta \leq 1/\sigma_{\max}^2$, they are all non-negative.

(e) Compute $(I - \eta X^\top X)^k$ in closed form for any $k \geq 1$ based on V, η and Λ.

Solution: The blessing of having the eigendecomposition shows up here: we have

$$(I - \eta X^\top X)^k = V(I - \eta\Lambda)^k V^\top.$$

We now compute parts (A) and (B) in Equation (30) separately. *From now on, we assume that* $\eta \leq 1/\sigma_{\max}^2$.

(f) If $\boldsymbol{v}^i$ is an eigenvector of $X^\top X$ corresponding to the eigenvalue σ_i^2, compute $(I - \eta X^\top X)^k \boldsymbol{v}^i$. Describe what happens when $k \to \infty$. (*Hint: you have to consider two cases: when* $\sigma_i = 0$ *and* $\sigma_i > 0$)

Solution: Note that the eigenvectors of $X^\top X$ and $I - \eta X^\top X$ are the same. Hence, if $\boldsymbol{v}^i$ is an eigenvector for $X^\top X$ corresponding to eigenvalue σ_i^2, it is also an eigenvector for $I - \eta X^\top X$ with eigenvalue $(1 - \eta\sigma_i^2)$. Hence,

$$(I - \eta X^\top X)^k \boldsymbol{v}^i = (1 - \eta\sigma_i^2)^k \boldsymbol{v}^i.$$

If $\sigma_i = 0$, then $(I - \eta X^\top X)^k \boldsymbol{v}^i = \boldsymbol{v}^i$ for any k.

If $\sigma_i > 0$, then $1 - \eta\sigma_i^2 < 1$ and as $k \to \infty$, $(1 - \eta\sigma_i^2)^k \to 0$ and hence $(I - \eta X^\top X)^k \boldsymbol{v}^i \to \boldsymbol{0}$.

(g) Based on the last step, compute part (A) when $k \to \infty$. (*Hint: decompose* $\boldsymbol{w}^0 = \boldsymbol{v} + \boldsymbol{u}$, *where* $\boldsymbol{u} \in \ker(X)$ *and* $\boldsymbol{v} \in \ker(X)^\perp$)

Solution: It is a fact from linear algebra that the right singular vectors (columns of V) corresponding to zero singular values form a basis of the kernel of the matrix.

In our case, this means that $\sigma_i = 0$ only if $\boldsymbol{v}^i \in \ker(X)$. Thus, if we decompose $\boldsymbol{w}^0 = \boldsymbol{v} + \boldsymbol{u}$ as in the hint, we have that $\boldsymbol{u} = \sum_{i:\sigma_i=0} \alpha_i \boldsymbol{v}^i$ (since $\boldsymbol{u} \in \ker(X)$ and $\boldsymbol{v}^i$ form a basis for the kernel), and $\boldsymbol{v} = \sum_{j:\sigma_j>0} \beta_j \boldsymbol{v}^j$. We can thus write

$$(I - \eta X^\top X)^k \boldsymbol{w}^0 = \sum_{i:\sigma_i=0} \alpha_i (I - \eta X^\top X)^k \boldsymbol{v}^i + \sum_{j:\sigma_j>0} \underbrace{\beta_j (I - \eta X^\top X)^k \boldsymbol{v}^j}_{\to 0 \quad \text{as} \quad k\to\infty}$$

(h) Show that part (B) is equal to

$$V\left(\eta \sum_{j=0}^{k-1} (I - \eta\Lambda)^j\right)\Sigma^\top U^\top \boldsymbol{y}$$

Solution:

$$\eta\left(\sum_{j=0}^{k-1} (I - \eta X^\top X)^j X^\top \boldsymbol{y}\right) = \eta\left(\sum_{j=0}^{k-1} V(I - \eta\Lambda)^j V^\top\right) V\Sigma^\top U^\top \boldsymbol{y} = V\left(\eta \sum_{j=0}^{k-1} (I - \eta\Lambda)^j\right)\Sigma^\top U^\top \boldsymbol{y}$$

(i) Compute (B) when $k \to \infty$. (*Hint: treat zero and positive singular values separately.*)

Solution: For simplicity, define

$$M := \sum_{j=0}^{k-1} (I - \eta\Lambda)^j.$$

Note that M is a diagonal $d \times d$ matrix. Its ith entry on the diagonal is

$$1 + (1 - \eta\sigma_i^2) + \cdots + (1 - \eta\sigma_i^2)^{k-1} = \begin{cases} k & \text{if} \quad \sigma_i = 0 \\ \frac{1-(1-\eta\sigma_i^2)^k}{\eta\sigma_i^2} & \text{if} \quad 0 < \sigma_i < 1/\sqrt{\eta} \end{cases}$$

Now note that $M\Sigma^\top$ is rectangular diagonal, and its ith element on the diagonal will be

$$M_{ii}\Sigma_{ii} = \begin{cases} 0 & \text{if} \quad \sigma_i = 0 \\ \frac{1-(1-\eta\sigma_i^2)^k}{\eta\sigma_i} & \text{if} \quad 0 < \sigma_i < 1/\sqrt{\eta} \end{cases}$$

Hence, when $k \to \infty$, the matrix $\eta M\Sigma^\top$ converges to the rectangular diagonal matrix N with diagonal entries

$$N_{ii} = \begin{cases} 0 & \text{if} \quad \sigma_i = 0 \\ \frac{1}{\sigma_i} & \text{if} \quad \sigma_i > 0 \end{cases}.$$

(j) Prove that the limit computed above equals $X^+\boldsymbol{y}$, where X^+ is the Moore-Penrose pseudo-inverse.

Solution: It is known that $VNU^\top$ is indeed the Moore-Penrose pseudo-inverse. Hence, part (B) equals $X^+\boldsymbol{y}$ when $k \to \infty$. Further, it can be proved that the Moore-Penrose pseudo-inverse provides the solution $w = X^+y$ with minimal Euclidean norm $\|w\|$ among all possible solutions.

(k) Notice that the above argument also works for the case where $n \geq d$. Make the necessary adjustments and prove that gradient descent initialized at $\mathbf{0}$ and with a small enough step size converges to the correct solution in under-parameterized setting.

Solution: The only thing that changes is the dimensions of Λ. The rest of the proof follows.