


Homework 1 (Linear Regression, Optimization)

For questions, please refer to Moodle.
Released on **1 March, 2022**

GENERAL INSTRUCTIONS

- This week's exercise is longer than usual to bridge the week of no lecture, deepen your knowledge of the fundamentals taught in the first two weeks, and remind yourself of some basic mathematical concepts used in the course.
- Submission of solutions is not mandatory but solving the exercises are highly recommended. The master solution will be released next week.
- Part of the exercises are available on Moodle. These problems are marked with .


Exercise 1: Math recap: multivariate normal distribution

This exercise serves as a recap of concepts in probability you should be familiar with to take the course. Recall the following fact about characteristic functions: For a random vector \mathbf{X} in \mathbb{R}^d , define its characteristic function $\varphi_{\mathbf{X}}$ as

$$\varphi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}[\exp(i\mathbf{t}^\top \mathbf{X})], \quad \text{for all } \mathbf{t} \in \mathbb{R}^d.$$

The characteristic function completely identifies a distribution. For a multivariate Normal distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, its characteristic function can be computed explicitly:

$$\varphi(\mathbf{t}) = \exp(i\mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top \Sigma \mathbf{t}).$$


- (a)  Let $\mathbf{X} = (X_1, \dots, X_d)$ be a d -dimensional standard Gaussian random vector, that is, $\mathbf{X} \sim \mathcal{N}_d(\mathbf{0}, I)$. Define $\mathbf{Y} = A\mathbf{X} + \boldsymbol{\mu}$, where A is a $d \times d$ matrix and $\boldsymbol{\mu} \in \mathbb{R}^d$. What is the distribution of \mathbf{Y} ?

Solution: Let us compute the characteristic function of \mathbf{Y} . Define $\mathbf{s} = A^\top \mathbf{t}$. We have

$$\begin{aligned} \varphi_{\mathbf{Y}}(\mathbf{t}) &= \mathbb{E}[\exp(i\mathbf{t}^\top \mathbf{Y})] \\ &= \mathbb{E}[\exp(i\mathbf{t}^\top A\mathbf{X}) \cdot \exp(i\mathbf{t}^\top \boldsymbol{\mu})] \\ &= \mathbb{E}[\exp(i\mathbf{s}^\top \mathbf{X})] \cdot \exp(i\mathbf{t}^\top \boldsymbol{\mu}) \\ &= \varphi_{\mathbf{X}}(\mathbf{s}) \cdot \exp(i\mathbf{t}^\top \boldsymbol{\mu}) \\ &= \exp(-\frac{1}{2}\mathbf{s}^\top \mathbf{s} + i\mathbf{t}^\top \boldsymbol{\mu}) \\ &= \exp(i\mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top A A^\top \mathbf{t}), \end{aligned}$$

which means that $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, A A^\top)$.

$$1. \mathcal{N}(\boldsymbol{\mu}, A) \quad 2. \mathcal{N}(\boldsymbol{\mu}, A^\top A) \quad 3. \mathcal{N}(\boldsymbol{\mu}, A^2) \quad 4. \mathcal{N}(\boldsymbol{\mu}, A A^\top)$$

- (b)  If B is an $r \times d$ matrix, what is the distribution of $B\mathbf{Y}$?

$$1. \mathcal{N}(B\boldsymbol{\mu}, B A A^\top B^\top) \quad 2. \mathcal{N}(B\boldsymbol{\mu}, B A A^\top) \quad 3. \mathcal{N}(\boldsymbol{\mu}, B A A^\top B^\top) \quad 4. \mathcal{N}(\boldsymbol{\mu}, B A A^\top)$$

Solution: With the same argument as the previous question, one gets $BY \sim \mathcal{N}(B\mu, BAA^\top B^\top)$.

- (c) [✓] Let $X = (X_1, X_2)$ be a bivariate Normal random variable with mean $\mu = (1, 1)$ and covariance matrix $\Sigma = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$. What is the mean and the variance of the conditional distribution of $Y = X_1 + X_2$ given $Z = X_1 - X_2 = 0$?

Solution: The correct answer mean 2, variance $\frac{20}{3}$

First, take a look at the following facts:

Let A, B be events. The definition of conditional probability $\mathbb{P}(A | B)$ assumes that $\mathbb{P}(B) \neq 0$. So one essentially cannot condition on events of zero probability in the usual way. The following is a workaround to this issue.

Let X, Y be random variables with joint density f and joint CDF F . For $\varepsilon > 0$ and $x, y \in \mathbb{R}$, we compute

$$\begin{aligned} \mathbb{P}(X \leq x | Y \in [y, y + \varepsilon]) &= \frac{\mathbb{P}(X \leq x, Y \in [y, y + \varepsilon])}{\mathbb{P}(Y \in [y, y + \varepsilon])} \\ &= \frac{F(x, y + \varepsilon) - F(x, y)}{F_Y(y + \varepsilon) - F_Y(y)} \\ &= \frac{[F(x, y + \varepsilon) - F(x, y)] / \varepsilon}{[F_Y(y + \varepsilon) - F_Y(y)] / \varepsilon}. \end{aligned}$$

Now if $\varepsilon \rightarrow 0$, the right-hand side has the limit $\frac{\partial_y F(x, y)}{f_Y(y)}$, and the left-hand side can be regarded as $\mathbb{P}(X \leq x | Y = y)$. Taking derivative with respect to x gives the conditional density

$$f_{X|Y}(x | y) = \frac{f(x, y)}{f_Y(y)}.$$

One can use this density to compute probabilities like $\mathbb{P}(X \in A | Y = y) = \iint_A \frac{f(x, y)}{f_Y(y)} dx dy$.

We present two approaches for this exercise:

APPROACH 1. Note that $Z = 0$ implies $X_1 = X_2$. Furthermore, by the definition of Y , we have $X_1 = X_2 = Y/2$ given $Z = 0$. Hence, the marginal density of Y given $Z = 0$ is proportional to

$$f_{Y|Z}(y | 0) = \frac{f_{Y,Z}(y, 0)}{f_Z(0)} \propto f_{Y,Z}(y, 0) \propto f_X \left[\begin{pmatrix} y/2 \\ y/2 \end{pmatrix} \right].$$

The last equality is due to the fact that the linear map $(x_1, x_2) \mapsto (x_1 + x_2, x_1 - x_2)$ has constant determinant of -2 . Thus, by a change of variables formula, the density changes by a constant factor. We then have

$$\begin{aligned} f_X \left[\begin{pmatrix} y/2 \\ y/2 \end{pmatrix} \right] &\propto \exp \left(-\frac{1}{2} \begin{pmatrix} \frac{y}{2} - 1 \\ \frac{y}{2} - 1 \end{pmatrix}^T \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}^{-1} \begin{pmatrix} \frac{y}{2} - 1 \\ \frac{y}{2} - 1 \end{pmatrix} \right) \\ &= \exp \left(-\frac{1}{2} \begin{pmatrix} \frac{y}{2} - 1 \\ \frac{y}{2} - 1 \end{pmatrix}^T \frac{1}{5} \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} \frac{y}{2} - 1 \\ \frac{y}{2} - 1 \end{pmatrix} \right) \\ &= \exp \left(-\frac{1}{2} \frac{(y - 2)^2}{\frac{20}{3}} \right). \end{aligned}$$

Clearly, the conditional distribution of Y given $Z = 0$ is hence Normal with mean 2 and variance $\frac{20}{3}$.

In this problem, we used the following trick which prevents a lot of computational headaches. If one is trying to derive the density of a random variable X at x , that is, $f_X(x)$, it is easier to neglect all *multiplicative* terms that does not include x . The reason is simply because $\int_{\mathbb{R}} f_X(x) dx = 1$.

Two important examples are single variable Normal random variables and multivariate Gaussian vectors. In the first case, following the trick above, we conclude that if a density function is of the form

$$f(x) \propto \exp(-ax^2 + bx)$$

for $a > 0$ and $b \in \mathbb{R}$, by completing the squares, we obtain

$$-ax^2 + bx = -a\left(x - \frac{b}{2a}\right)^2 + \frac{b^2}{4a},$$

and thus, by removing the terms that do not depend on x , we get

$$f(x) \propto \exp\left(-\frac{1}{2} \frac{\left(x - \frac{b}{2a}\right)^2}{1/2a}\right),$$

meaning that the distribution is a normal distribution with mean $\frac{b}{2a}$ and variance $1/2a$.

The situation for multivariate normal distribution is the same. One needs only to create a proper quadratic form in the exponent to get the familiar multivariate Gaussian density.

APPROACH 2. We define the random vector \mathbf{R} as

$$\mathbf{R} = \begin{pmatrix} Y \\ Z \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}}_{=A} \mathbf{X}.$$

Notice that \mathbf{R} is a linear transformation of a Gaussian vector, and by part (a), it is a Gaussian vector. Thus, we only need to compute its mean and covariance matrix. By linearity of expectation, the mean $\mu_{\mathbf{R}}$ of \mathbf{R} is

$$\mathbb{E}[\mathbf{R}] = A \mathbb{E}[\mathbf{X}] = A\boldsymbol{\mu} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}.$$

The covariance matrix $\Sigma_{\mathbf{R}}$ of \mathbf{R} is also given by part (a):

$$\Sigma_{\mathbf{R}} = A \Sigma A^{\top} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 7 & 1 \\ 1 & 3 \end{pmatrix}$$

The conditional density of Y given $Z = 0$ is then given by

$$\begin{aligned} f_{Y|Z}(y | 0) &= \frac{f_{Y,Z}(y, 0)}{f_Z(0)} \propto f_{Y,Z}(y, 0) \\ &\propto \exp\left(-\frac{1}{2} \begin{pmatrix} y-2 \\ 0 \end{pmatrix}^{\top} \begin{pmatrix} 7 & 1 \\ 1 & 3 \end{pmatrix}^{-1} \begin{pmatrix} y-2 \\ 0 \end{pmatrix}\right) \\ &= \exp\left(-\frac{1}{2} \begin{pmatrix} y-2 \\ 0 \end{pmatrix}^{\top} \frac{1}{20} \begin{pmatrix} 3 & -1 \\ -1 & 7 \end{pmatrix} \begin{pmatrix} y-2 \\ 0 \end{pmatrix}\right) \\ &= \exp\left(-\frac{1}{2} \frac{(y-2)^2}{\frac{20}{3}}\right). \end{aligned}$$

Clearly, the conditional distribution of Y given $Z = 0$ is hence Normal with mean 2 and variance $\frac{20}{3}$.

- (d) For $M \sim \mathcal{N}_n(\mathbf{0}, I)$, we say that the random variable $V = \|M\|^2$ has the χ^2 (chi-square) distribution with n degrees of freedom, written as $V \sim \chi^2(d)$. Assume that X_1, \dots, X_n are i.i.d. samples from the Normal distribution $\mathcal{N}(\mu, \sigma^2)$. One way to estimate σ^2 from these samples is to look at the sample variance:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2,$$

where $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$. Note that S^2 is computed based on X_1, \dots, X_n , hence it is a random variable. By following the steps below, prove that $\frac{(n-1)}{\sigma^2} S^2$ has a chi-square distribution with $n-1$ degrees of freedom.

Step 1. Prove that the random vector $\mathbf{Y} = (X_1 - \bar{X}, \dots, X_n - \bar{X})$ satisfies $\|\mathbf{Y}\|^2 = (n-1)S^2$.

Step 2. Recall and verify the following facts about projections:

Let \mathbf{v} be a unit vector in \mathbb{R}^d . The orthogonal projection on the hyperplane defined by \mathbf{v} is then $I - \mathbf{v}\mathbf{v}^\top$. Also, the reflection about the hyperplane defined by \mathbf{v} is $I - 2\mathbf{v}\mathbf{v}^\top$ (verify these by drawing a picture). Sometimes, the last transformation is called a Householder Reflector. For given vectors \mathbf{u}, \mathbf{v} , if one is searching for an orthogonal matrix that maps \mathbf{u} to \mathbf{v} , one possible way is to consider the Householder reflector about the hyperplane defined by $(\mathbf{v} - \mathbf{u})/\|\mathbf{v} - \mathbf{u}\|$.

- Step 3. Define $\mathbf{X} = (X_1, \dots, X_n)$. Construct the vector \mathbf{v} such that \mathbf{Y} is the orthogonal projection of \mathbf{X} on the hyperplane defined by \mathbf{v} . Conclude that \mathbf{Y} has a Gaussian distribution.
- Step 4. As projection on a hyperplane is an operator with rank $n-1$, it is more convenient to transform \mathbf{Y} in a way that one of its components becomes zero, while its norm is kept fixed (that is, we are looking for an orthogonal transformation). Show that the Householder reflector (defined above) that takes \mathbf{v} to $\mathbf{w} := (1, 0, \dots, 0)$ is such a transformation.
- Step 5. Let \mathbf{Z} be the reflection of \mathbf{Y} . Prove that it has a Gaussian distribution, find its mean and covariance matrix, and conclude the problem.

Solution: The fact that X_i are i.i.d. implies that $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$, where $\boldsymbol{\mu} = (\mu, \dots, \mu)$. Consider the unit vector $\mathbf{v} = (1/\sqrt{n}, \dots, 1/\sqrt{n})$. The projection of \mathbf{X} on the direction of \mathbf{v} is

$$\mathbf{v}\mathbf{v}^\top \mathbf{X} = \begin{pmatrix} 1/n & \cdots & 1/n \\ \vdots & \ddots & \vdots \\ 1/n & \cdots & 1/n \end{pmatrix} \mathbf{X} = \begin{pmatrix} \bar{X} \\ \vdots \\ \bar{X} \end{pmatrix}.$$

Thus, the projection on the hyperplane defined by \mathbf{v} is the vector $\mathbf{Y} = (I - \mathbf{v}\mathbf{v}^\top)\mathbf{X} = (X_1 - \bar{X}, \dots, X_n - \bar{X})$. Observe that \mathbf{Y} is a Gaussian vector, as it is a linear function of \mathbf{X} .

Using Householder reflectors, the reflection map is indeed $I - 2\mathbf{u}\mathbf{u}^\top$, where $\mathbf{u} = (\mathbf{v} - \mathbf{w})/\|\mathbf{v} - \mathbf{w}\|$.

Denote by $\mathbf{Z} = (I - 2\mathbf{u}\mathbf{u}^\top)\mathbf{Y}$. Observe again that \mathbf{Z} is a Gaussian vector. It is easy to verify that the mean of \mathbf{Z} is zero. The covariance matrix can be computed using part (a) and (b):


$$\begin{aligned} \Sigma_{\mathbf{Z}} &= (I - 2\mathbf{u}\mathbf{u}^\top)(I - \mathbf{v}\mathbf{v}^\top)(\sigma^2 I)(I - \mathbf{v}\mathbf{v}^\top)^\top(I - 2\mathbf{u}\mathbf{u}^\top)^\top \\ &= \sigma^2(I - 2\mathbf{u}\mathbf{u}^\top)(I - \mathbf{v}\mathbf{v}^\top)(I - 2\mathbf{u}\mathbf{u}^\top) \\ &= \sigma^2 \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \end{aligned}$$

Thus, \mathbf{Z}/σ is a Gaussian random vector, that is supported on an $(n-1)$ -dimensional space, with mean 0 and covariance I_{n-1} . That is, it is a standard Gaussian vector in \mathbb{R}^{n-1} . Hence, $\frac{1}{\sigma^2}\|\mathbf{Z}\|^2$ has chi-square distribution with $(n-1)$ degrees of freedom. But $(n-1)S^2 = \|\mathbf{Y}\|^2 = \|\mathbf{Z}\|^2$. Thus,

$$\frac{(n-1)}{\sigma^2} S^2 \sim \chi^2(n-1).$$

Exercise 2: Local vs. Global Optima

To decide whether the statement is True or False, prove it or provide a counter-example.

- (a)  In this exercise, you learn what type of optima (local, global, unique global) exist for convex and strictly convex functions.

A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is *convex* if $\text{dom } f$ is a convex set and if for all $x, y \in \text{dom } f$, and θ with $0 \leq \theta \leq 1$, we have

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$$

A function f is *strictly convex* if strict inequality holds whenever $x \neq y$ and $0 < \theta < 1$.

The sum of strictly convex and convex functions is a strictly convex function.

☒ True ☐ False

Solution: f is convex, g is strictly convex. $0 \leq \theta \leq 1$

$$\begin{aligned} (f + g)(\theta x + (1 - \theta)y) &= f(\theta x + (1 - \theta)y) + g(\theta x + (1 - \theta)y) \\ &< \theta f(x) + (1 - \theta)f(y) + \theta g(x) + (1 - \theta)g(y) \\ &= \theta(f + g)(x) + (1 - \theta)(f + g)(y) \end{aligned} \quad (1)$$

Any local minimum of a convex function is also a global minimum.

☒ True ☐ False

Solution: Let's prove the result by contradiction. Let x_{loc} be a local minimum and x_{glob} a global minimum such that $f(x_{\text{glob}}) < f(x_{\text{loc}})$. Since x_{loc} is a local minimum, there exists $\alpha > 0$ for which $f(x_{\text{loc}}) \leq f(x)$ for all $x \in \mathbb{R}^n$ such that $\|x - x_{\text{loc}}\|_2 \leq \alpha$. Let's choose $\theta \in (0, 1)$ such that $x_\theta := \theta x_{\text{loc}} + (1 - \theta)x_{\text{glob}}$ satisfies $\|x_\theta - x_{\text{loc}}\|_2 \leq \alpha$. Therefore

$$\begin{aligned} f(x_{\text{loc}}) &\leq f(x_\theta) \\ &\leq \theta f(x_{\text{loc}}) + (1 - \theta)f(x_{\text{glob}}) \quad \text{by convexity of } f \\ &< f(x_{\text{loc}}) \quad \text{because } f(x_{\text{glob}}) < f(x_{\text{loc}}) \end{aligned}$$

Every strictly convex function has a positive definite Hessian and a unique global minimum.

☐ True ☒ False

Solution: The positive definiteness of $f(x)$ is sufficient but not necessary for the strict convexity. Counter-example $f(x) = x^4$.

However, if the Hessian is positive definite, the function is strictly convex and has a unique global minimum.

- (b)  Here we look at the properties of the Hessian and how they relate to optimality.

The Hessian of non-convex function is negative semi-definite.

☐ True ☒ False

Solution: If the function is non-convex, it doesn't mean it is a concave: $f(x) = \sin(x)$. Concave function has a negative semi-definite Hessian.

If the Hessian of a convex function $f(x)$ is positive-definite at x_0 , x_0 is a local minimum.

☐ True ☒ False

Solution: The first order derivative must also be 0 at point x_0 . Otherwise, every x would be a minimum for $f(x) = x^2$.

The function f has a local minimum at point x_0 if its gradient equals 0 at x_0 , i.e., $\nabla f(x_0) = 0$ and the determinant of the Hessian matrix is positive at x_0 .

☐ True ☒ False

Solution: It would be True if the Hessian is positive semi-definite.

Consider the second order Taylor expansion around critical point x_0 .

$$f(x) = f(x_0) + \nabla f(x_0)(x - x_0) + \frac{1}{2}(x - x_0)^T H(x - x_0)$$

at a critical point the gradient vanishes $\nabla f(x_0) = 0$


$$f(x) = f(x_0) + \frac{1}{2}(x - x_0)^T H(x - x_0)$$

For a minimum, one would need

$$f(x) - f(x_0) = \frac{1}{2}(x - x_0)^T H(x - x_0) \geq 0$$

and that is why positive semi-definiteness is needed.

Criterion: the Hessian is positive-definite iff the determinant of all upper-left submatrices ≥ 0 .

(c)  Practice applying the concepts you have learned in the following examples.

The set of all orthogonal $n \times n$ matrices is a convex set in $\mathbb{R}^{n \times n}$.

☐ True ☒ False

Solution: Take $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. They are both orthogonal but their midpoint $(A + B)/2 = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$ is definitely not orthogonal.

$f(x_1, x_2) = 1/(x_1 x_2)$ on \mathbb{R}_{++}^2 (all non-negative real numbers) is convex.

☒ True ☐ False

Solution: The Hessian of f is

$$\nabla^2 f(x) = \frac{1}{x_1 x_2} \begin{bmatrix} 2/(x_1^2) & 1/(x_1 x_2) \\ 1/(x_1 x_2) & 2/x_2^2 \end{bmatrix} \succ 0$$

Consider the following function

$$f(\mathbf{w}) = \|\Phi\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|^2$$

□ True ■ False

where $\lambda > 0$, matrix $\Phi \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \mathbb{R}^n$, and $\mathbf{w} \in \mathbb{R}^d$. Is it true that f has a unique global minimum if and only if $n \geq d$ and the columns of Φ are independent?

Solution: In order to show that global minimum exists and is unique, we prove that Hessian is positive definite. We also show that it holds for any Φ .

We start with computing the derivative

$$Df(\mathbf{w}) = 2(\Phi\mathbf{w} - \mathbf{y})^\top \Phi + 2\lambda\mathbf{w}^\top = 2\mathbf{w}^\top (\Phi^\top \Phi + \lambda I_d) - 2\mathbf{y}^\top \Phi$$

The Hessian is the derivative of the derivative, which is equal to

$$D^2f(\mathbf{w}) = 2(\Phi^\top \Phi + \lambda I_d).$$

To prove that the Hessian matrix is positive-definite, we show that for any non-zero vector $\mathbf{v} \in \mathbb{R}^d$, we have $\mathbf{v}^\top D^2f(\mathbf{w})\mathbf{v} > 0$. Indeed,

$$\begin{aligned} \mathbf{v}^\top D^2f(\mathbf{w})\mathbf{v} &= 2\mathbf{v}^\top (\Phi^\top \Phi + \lambda I_d)\mathbf{v} \\ &= 2\mathbf{v}^\top \Phi^\top \Phi \mathbf{v} + 2\lambda\mathbf{v}^\top I_d \mathbf{v} \\ &= 2\|\Phi\mathbf{v}\|^2 + 2\lambda\|\mathbf{v}\|^2 > 0, \end{aligned}$$

This holds for any matrix Φ simply because $\lambda > 0$.

Positive-definite Hessian implies that the function is strictly convex. Strictly convex function has a unique global minimum.

Note: this function defines *Ridge regression*.

Exercise 3: 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 $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R} \times \mathbb{R}$ and the hypothesis space of affine functions $H = \{w_0 + w_1x : w_0, w_1 \in \mathbb{R}\}$. The error of a solution $f \in H$, i.e., $f(x) = w_0 + w_1x$ for some $w_0, w_1 \in \mathbb{R}$ is given by

$$L(w_0, w_1) = \frac{1}{2n} \sum_{i=1}^n (y_i - w_0 - w_1x_i)^2. \quad (2)$$

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

$$\frac{\partial L(0, w_1)}{\partial w_1} = \frac{1}{n} \sum_{i=1}^n (w_1x_i - y_i)x_i \quad (3)$$

and the second derivative

$$\frac{\partial^2 L(0, w_1)}{\partial w_1^2} = \frac{1}{n} \sum_{i=1}^n x_i^2. \quad (4)$$

We observe that the second derivative (4) 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 (3) to zero

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (w_1 x_i - y_i) x_i &= 0 \\ \Leftrightarrow w_1 &= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i}. \end{aligned} \quad (5)$$

If not all x_i are equal to zero w_1^* is given by (5).

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

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

Let

$$\Phi = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}. \quad (6)$$

Now we can write (2) using matrix notation $L(w) = \frac{1}{2n} \|y - \Phi w\|^2$, where $w = (w_0, w_1)^T$.

The gradient of (2) is

$$\nabla L(w) = \frac{1}{n} \Phi^T (\Phi w - y) \quad (7)$$

and the Hessian is

$$\nabla^2 L(w) = \frac{1}{n} \Phi^T \Phi. \quad (8)$$

The matrix $\Phi^T \Phi$ is positive definite if it satisfies $v^T \Phi^T \Phi v > 0$ for all $v \neq 0 \in \mathbb{R}^2$. We observe that $v^T \Phi^T \Phi v > 0$ is equivalent to $\|\Phi v\|^2 > 0$, which in turn is equivalent to $\Phi v \neq 0$.

We are going to prove by contradiction that $\Phi v \neq 0$. Suppose that $\Phi v = 0$. Then we have

$$\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} v_1 + \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} v_2 = 0, \quad (9)$$

i.e., for all $i \in \{1, \dots, n\}$ we have $x_i = -\frac{v_1}{v_2}$ which contradicts the assumption $x_i \neq x_j$ 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

$$\nabla L(w_0, w_1) = \begin{pmatrix} \frac{\partial L(w_0, w_1)}{\partial w_0} \\ \frac{\partial L(w_0, w_1)}{\partial w_1} \end{pmatrix}. \quad (10)$$

Solution: We get

$$\nabla L(w_0, w_1) = \begin{pmatrix} \frac{\partial L(w_0, w_1)}{\partial w_0} \\ \frac{\partial L(w_0, w_1)}{\partial w_1} \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n w_0 + w_1 x_i - y_i \\ \frac{1}{n} \sum_{i=1}^n (w_0 + w_1 x_i - y_i) x_i \end{pmatrix}. \quad (11)$$

- (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 (10) 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

$$w_0^* = \frac{\sum_{i=1}^n y_i - w_1 \sum_{i=1}^n x_i}{n} = \bar{y} - w_1 \bar{x}, \quad (12)$$

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

$$w_1^* = \frac{\sum_{i=1}^n x_i y_i - x_i \bar{y}}{\sum_{i=1}^n x_i^2 - x_i \bar{x}}, \quad (13)$$

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

$$\begin{aligned} w_1^* &= \frac{\sum_{i=1}^n (x_i y_i - x_i \bar{y}) + \overbrace{\sum_{i=1}^n (\bar{x} \bar{y} - \bar{x} y_i)}^{=0}}{\sum_{i=1}^n (x_i^2 - x_i \bar{x}) + \underbrace{\sum_{i=1}^n (\bar{x}^2 - \bar{x} x_i)}_{=0}} \\ &= \frac{\sum_{i=1}^n (\bar{x} - x_i)(\bar{y} - y_i)}{\sum_{i=1}^n (\bar{x} - x_i)^2} \\ &= \frac{\text{Cov}(x, y)}{\text{Var}(x)}. \end{aligned} \quad (14)$$

Another way of writing (2) is in matrix notation

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

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

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

is the data matrix.

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

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

to zero.

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

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

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: From linear algebra we know that $\Phi^T \Phi \in \mathbb{R}^{(d+1) \times (d+1)}$ is invertible if it has full rank.

For real matrices $\Phi \in \mathbb{R}^{n \times (d+1)}$ we have $\text{rank}(\Phi) = \text{rank}(\Phi^T \Phi)$. This follows from the rank-nullity theorem together with the fact that Φ and $\Phi^T \Phi$ have the same null spaces.

Recall the rank-nullity theorem from linear algebra which states that for a linear mapping $T : V \rightarrow W$ we have $\text{rank}(T) + \text{nullity}(T) = \dim(V)$, where $\text{rank}(T) = \dim(\text{img}(T))$, $\text{nullity}(T) = \dim(\ker(T))$, \dim is the dimension of a vector space, img is the image of a map and \ker is the kernel of T , i.e., the set of vectors that satisfy $T(v) = 0$. As there is an isomorphism between matrices and linear mappings we get also a rank-nullity theorem for matrices.

Now, more formally for real matrices $\Phi \in \mathbb{R}^{n \times (d+1)}$ we have $\text{rank}(\Phi) = \text{rank}(\Phi^T \Phi)$, because

$$\ker(\Phi^T \Phi) = \{w : \Phi^T \Phi w = 0\} \quad (19)$$

is equal to

$$\ker(\Phi) = \{w : \Phi w = 0\}. \quad (20)$$

The equality of the null spaces follows from

$$\begin{aligned} \Phi w &= 0 \\ \Leftrightarrow \|\Phi w\|^2 &= 0 \\ \Leftrightarrow w^T \Phi^T \Phi w &= 0 \\ \Leftrightarrow \Phi^T \Phi w &= 0. \end{aligned} \quad (21)$$

As a consequence, we obtain the necessary condition: If $\text{rank}(\Phi) = d + 1$, $\Phi^T \Phi$ is invertible.

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

Solution: See the lecture for a complete proof. Here is yet another proof based on convexity. Note that $L : \mathbb{R}^p \rightarrow \mathbb{R}$ and

$$L(w) = \frac{1}{2n} \|\Phi w - y\|^2 = (\Phi w - y)^T (\Phi w - y) = w^T \Phi^T \Phi w - 2w^T \Phi^T y + y^T y.$$

The gradient of this function is equal to (see the second tutorial; also note that the gradient is a vector in \mathbb{R}^d)

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

Because $L(w)$ is convex (formally proven in (d)), its optima (if they exist) are exactly those points that have a zero gradient, i.e., those w^* that satisfy $\Phi^T \Phi w^* = \Phi^T y$. Under the given assumption, the unique minimizer is indeed equal to $w^* = (\Phi^T \Phi)^{-1} \Phi^T y$.

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

$$\min_{w \in \mathbb{R}^{d+1}} \|y - \Phi w\|^2 \quad (22)$$

does not admit a unique solution.

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

$$\dim(\ker(\Phi^T \Phi)) = d + 1 - \text{rank}(\Phi^T \Phi) = d + 1 - n > 0. \quad (23)$$

Our objective $\|y - \Phi w\|^2$ is convex (recall from the tutorial $\nabla^2 \|y - \Phi w\|^2 = 2\Phi^T \Phi \succeq 0$ for all w). Thus, each minimal weight vector $w^* \in \arg \min_w \|y - \Phi w\|^2$ satisfies $\nabla \|y - \Phi w^*\|^2 = 0 \Leftrightarrow \Phi^T \Phi w^* = \Phi^T y$. Now, given a minimal weight vector w^* , we have $\Phi^T \Phi(w^* + u) = \Phi^T y$ for all $u \in \ker(\Phi^T \Phi)$. As $\dim(\ker(\Phi^T \Phi)) \geq 1$ there are infinitely many such $u \neq 0$.

Remember the gradient descent update as discussed in lecture: $w^{t+1} = w^t - \eta \nabla L(w)$. In the following questions we would like to compare the computational complexity of the closed-form solution $w^* = (\Phi^T \Phi)^{-1} \Phi^T y$ against the one of the gradient descent algorithm. For the next exercises (h) and (i) you may use the contraction inequality as discussed during lecture,

$$\|w^{t+1} - w^*\|_2 \leq \|I - \eta \Phi^T \Phi\|_{op} \|w^t - w^*\|_2.$$

- (h) 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 w^τ that satisfies $\|w^\tau - w^*\| < \epsilon$, where w^τ is the parameter vector computed by gradient descent after τ steps.

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

$$\begin{aligned} \rho^\tau \|w^0 - w^*\| &< \epsilon \\ \Leftrightarrow \rho^\tau &< \frac{\epsilon}{\|w^0 - w^*\|} \\ \Leftrightarrow \tau \log \rho &< \log \epsilon - \log \|w^0 - w^*\| \\ \Leftrightarrow \tau &> \frac{\log \epsilon - \log \|w^0 - w^*\|}{\log \rho}, \end{aligned} \quad (24)$$

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 $\nabla L(w^t) = \frac{1}{n}(\Phi^T \Phi w^t - \Phi^T y)$, requires $O(np)$ operations to compute $\Phi^T y$, $O(p^2)$ operations to compute $\Phi^T \Phi w^t$ and one time $O(np^2)$ operations to compute $\Phi^T \Phi$ at the beginning of the algorithm. Subtracting $\eta \nabla L(w^t)$ 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 $\tau = \lceil \frac{\log \epsilon - \log \|w^0 - w^*\|}{\log \rho} \rceil$ steps, resulting in a computational complexity of $O\left(np^2 + \frac{\log \epsilon - \log \|w^0 - w^*\|}{\log \rho} (p^2 + np)\right)$.

- (i) Now, say you are free to choose a constant stepsize. What is the “minimum” number of iterations τ required to obtain a solution w^τ that satisfies $\|w^\tau - w^*\| < \epsilon$? 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

$$\rho = \|I - \eta \Phi^T \Phi\| = \max(|1 - \eta \lambda_{\max}|, |1 - \eta \lambda_{\min}|), \quad (25)$$

where λ_{\min} and λ_{\max} denote the smallest and the largest eigenvalue of $\Phi^T \Phi$. In the lecture we saw that $\eta < \frac{2}{\lambda_{\max}}$ is required for convergence, i.e., for $\rho < 1$.

Now, in order to determine the optimal choice of η we first derive an expression for ρ as a function of η . We have

$$\begin{aligned}\rho(\eta) &= \max(|1 - \eta\lambda_{\max}|, |1 - \eta\lambda_{\min}|) \\ &= \max(1 - \eta\lambda_{\max}, \eta\lambda_{\max} - 1, 1 - \eta\lambda_{\min}, \eta\lambda_{\min} - 1) \\ &= \max(\eta\lambda_{\max} - 1, 1 - \eta\lambda_{\min}),\end{aligned}\tag{26}$$

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

$$\rho(\eta) = \begin{cases} 1 - \eta\lambda_{\min} & \text{for } 0 < \eta < \frac{2}{\lambda_{\min} + \lambda_{\max}}, \\ \eta\lambda_{\max} - 1 & \text{for } \frac{2}{\lambda_{\min} + \lambda_{\max}} \leq \eta < \frac{1}{\lambda_{\max}}, \end{cases}\tag{27}$$

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

$$\begin{aligned}\rho(\eta^*) &= 1 - \frac{2\lambda_{\min}}{\lambda_{\min} + \lambda_{\max}} \\ &= \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\min} + \lambda_{\max}}.\end{aligned}\tag{28}$$

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

$$\tau = \lceil \frac{\log \epsilon - \log \|w^0 - w^*\|}{\log(\lambda_{\max} - \lambda_{\min}) - \log(\lambda_{\min} + \lambda_{\max})} \rceil.\tag{29}$$

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

Solution: The complexity of gradient descent is $O(np^2 + \tau(p^2 + np))$. In comparison, the complexity of computing w^* in closed form, which is dominated by the cost of the matrix inversion $(\Phi^T \Phi)^{-1}$ is in $O(p^3)$. 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 $O(np^2)$ cost of computing $\Phi^T \Phi$ as solving the resulting $p \times p$ linear system of equations is in $O(p^2)$ 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 $(\Phi^T \Phi)^{-1}$ shows that gradient descent can be preferable if p is significantly larger than n and if the optimal τ from (29) is smaller than p .

Furthermore, using matrix notation we can also use more general nonlinear hypothesis spaces by using nonlinear basis functions $\phi_1 : \mathbb{R}^d \rightarrow \mathbb{R}, \dots, \phi_p : \mathbb{R}^d \rightarrow \mathbb{R}$ (see bonus exercise 6)

$$\Phi = \begin{pmatrix} \phi_1(x_1) & \phi_2(x_1) & \dots & \phi_p(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \dots & \phi_p(x_2) \\ \vdots & \vdots & & \vdots \\ \phi_1(x_n) & \phi_2(x_n) & \dots & \phi_p(x_n) \end{pmatrix}.\tag{30}$$

Now, fitting a nonlinear function to the data simply amounts to choosing basis functions ϕ_1, \dots, ϕ_p and solving the linear system of equations that is obtained by setting the gradient in (17) to zero.

Questions

- (j) For $x_1, \dots, x_n \in \mathbb{R}^d$, define basis functions ϕ_1, \dots, ϕ_p such that (30) specializes to (18).

Solution: We define the $p = d + 1$ basis functions $\phi_1(x) = 1, \phi_2(x) = x, \dots, \phi_{d+1}(x) = x_d$.

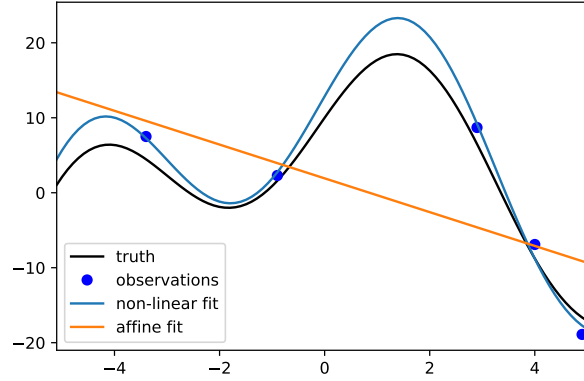


Figure 1: Linear regression with basis functions $\phi_1(x) = 1, \phi_2(x) = x, \phi_3(x) = x^2, \phi_4(x) = \sin(x)$. The true function underlying the data generating process is depicted in black. We use it to obtain noisy measurements, which are depicted as blue dots. Then, we fit an affine function and a non-linear function from the hypothesis space $H = \{w_1 + w_2x + w_3x^2 + w_4\sin(x) : w_1, w_2, w_3, w_4 \in \mathbb{R}\}$ to the measurements.

- (k) [✓] In Figure 1 we consider the data generating process $y = f(x) + \epsilon$ with $f(x) = 10 - 0.7x^2 + 10\sin(x)$ and $\epsilon \sim N(0, 2)$. Using this data generation process we generated a training set by sampling five measurement locations $x_1, \dots, x_5 \sim U([-5, 5])$ uniformly at random and then sampling their corresponding noisy function evaluations y_1, \dots, y_5 with $y_i \sim N(f(x_i), 2)$. As a result, we obtain the training set

$$\{(4, -6.9), (-0.9, 2.3), (2.9, 8.7), (-3.4, 7.5), (4.9, -18.9)\}.$$

Fit a nonlinear function to this dataset by solving the linear regression problem using $\phi_1(x) = 1, \phi_2(x) = x, \phi_3(x) = x^2, \phi_4(x) = \sin(x)$ as basis functions. Report the resulting coefficients $w_1^*, w_2^*, w_3^*, w_4^*$.

Solution: Evaluating our basis functions on the data points x_1, \dots, x_5 (rounded to two decimal places) yields

$$\Phi = \begin{pmatrix} 1 & 4 & 16 & -0.76 \\ 1 & -0.9 & 0.81 & -0.78 \\ 1 & 2.9 & 8.41 & 0.24 \\ 1 & -3.4 & 11.56 & 0.26 \\ 1 & 4.9 & 24.01 & -0.98 \end{pmatrix}. \quad (31)$$

Now, solving the linear system $\Phi^T \Phi w = \Phi^T y$ with $y = (-6.9, 2.3, 8.7, 7.5, -18.9)^T$ yields the solution (rounded to two decimal places) $w_1^* = 12.86, w_2^* = -0.05, w_3^* = -0.77, w_4^* = 12.19$.

Exercise 4: Gradient Descent for Linear Regression

In this exercise, we are going to prove that under mild conditions the gradient descent algorithm for ordinary linear regression problem 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 $\mathbf{y} \in \mathbb{R}^n$ is the response vector. The goal is to find a vector $\mathbf{w} \in \mathbb{R}^d$ such that $L(\mathbf{w}) := \frac{1}{2} \|X\mathbf{w} - \mathbf{y}\|^2$ is minimized. For this, we use the gradient descent algorithm: starting from an initial vector \mathbf{w}^0 , the iterates of the gradient descent algorithm for a step size η are

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

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

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

Solution: It is already done in the lecture

- (b) By using induction on k , prove that

$$\mathbf{w}^k = \underbrace{(I - \eta X^\top X)^k \mathbf{w}^0}_{(A)} + \underbrace{\eta \left(\sum_{j=0}^{k-1} (I - \eta X^\top X)^j \right) X^\top \mathbf{y}}_{(B)} \quad (32)$$

Solution: The base case ($k = 0$) is clear. Assuming that (32) holds for k , proving it for $k + 1$ amounts to plug in (32) into the gradient descent equation and verifying that it holds.

From (b) it is clear that powers of the matrix $I - \eta X^\top X$ play an important role in understanding what happens to \mathbf{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 $\sigma_1, \dots, \sigma_n$ 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 $I = V V^\top$ (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 is immediate that we have the eigen-decomposition 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 $\eta \leq 1/\sigma_{\max}^2$, all are 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 eigen-decomposition shows up here: we have

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

From now on, we assume that $\eta \leq 1/\sigma_{\max}^2$.

We now compute parts (A) and (B) in Eqn. (32) separately.

(f) If \mathbf{v}^i is an eigenvector of $X^\top X$ corresponding to the eigenvalue σ_i^2 , compute $(I - \eta X^\top X)^k \mathbf{v}^i$. Describe what happens when $k \rightarrow \infty$. (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 \mathbf{v}^i is an eigenvector for $X^\top X$ corresponding to eigenvalue σ_i^2 , it is also an eigenvector for $I - \eta X^\top X$ for the eigenvalue $(1 - \eta\sigma_i^2)$. Hence,

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

If $\sigma_i = 0$, then $(I - \eta X^\top X)^k \mathbf{v}^i = \mathbf{v}^i$ for any k , however, if $\sigma_i > 0$, then $1 - \eta\sigma_i^2 < 1$ and as $k \rightarrow \infty$, $(1 - \eta\sigma_i^2)^k \rightarrow 0$, which gives $(I - \eta X^\top X)^k \mathbf{v}^i \rightarrow \mathbf{0}$.

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

Solution: It is a fact from linear algebra that the right singular vectors corresponding to zero singular values form a basis of the kernel of the matrix. In our case, $\sigma_i = 0$ happens only if $\mathbf{v}^i \in \ker(X)$. Thus, if we decompose $\mathbf{w}^0 = \mathbf{v} + \mathbf{u}$ as in the hint, we have that $\mathbf{u} = \sum_{i:\sigma_i=0} \alpha_i \mathbf{v}^i$ (since $\mathbf{u} \in \ker(X)$ and \mathbf{v}^i form a basis for the kernel), and $\mathbf{v} = \sum_{j:\sigma_j>0} \beta_j \mathbf{v}^j$. We thus can write

$$\begin{aligned} (I - \eta X^\top X)^k \mathbf{w}^0 &= \sum_{i:\sigma_i=0} \alpha_i (I - \eta X^\top X)^k \mathbf{v}^i + \sum_{j:\sigma_j>0} \underbrace{\beta_j (I - \eta X^\top X)^k \mathbf{v}^j}_{\rightarrow 0 \text{ as } k \rightarrow \infty} \\ &\rightarrow \sum_{i:\sigma_i=0} \alpha_i \mathbf{v}^i = \mathbf{v}. \end{aligned}$$

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

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

Solution: It is obvious.

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

Solution: For simplicity, define

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

Note that A is a diagonal $d \times d$ matrix. Its i th entry on the diagonal is

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

Now note that $A\Sigma^\top$ is rectangular diagonal, and its i th element on the diagonal will be

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

Hence, when $k \rightarrow \infty$, the matrix $\eta A\Sigma^\top$ converges to the rectangular diagonal matrix B with diagonal entries

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

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

Solution: It is known that $V^\top B U^\top$ is indeed the Moore-Penrose pseudo-inverse. Hence, part (B) is nothing but X^+y .

- (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.

Exercise 5: Post your machine learning application

We created forums on [Moodle](#) for you to post an application that you find interesting and where you think machine learning could be a useful tool. You may want to address some of the following points if they are applicable:

- **Motivation:** Briefly describe why your application problem is interesting and needs to be solved.
 - **Problem description:** Describe your application as if you were describing it to a machine learning expert who would benefit from as many details in your description as possible
 - Which part exactly involves machine learning?
 - What is the learning task?
 - What is the data?
 - What are the inputs and outputs of the machine learning component?
 - How do you measure the quality of a solution?
- (a) Post a machine learning application in one of the Moodle forums or leave a comment (e.g., +1) on an interesting application posted by a fellow student.

Exercise 6: Bonus (not relevant for exam)

In this bonus exercise we will provide some mathematical background on the nonlinear hypothesis spaces that we used in exercise 3. In particular, we will show that they are linear vector spaces. Let's start by considering the hypothesis space H is a space of affine functions

$$f : \mathbb{R} \rightarrow \mathbb{R} : x \mapsto w_0 + w_1 x. \quad (33)$$

Recall from linear algebra that a linear vector space is a set V associated with an addition $\oplus : V \times V \rightarrow V$ and a scalar multiplication $\odot : \mathbb{R} \times V \rightarrow V$ that, for $a, b \in V$ and $\alpha, \beta \in \mathbb{R}$, satisfy

1. additive axioms:

(a) (V, \oplus) is a group,

2. multiplicative axioms:

(a) $0 \odot a = 0$,

(b) $1 \odot a = a$,

(c) $(\alpha\beta) \odot a = \alpha \odot (\beta \odot a)$,

3. and distributive axioms:

(a) $\alpha \odot (a \oplus b) = (\alpha \odot a) \oplus (\alpha \odot b)$,

(b) $(\alpha + \beta) \odot a = (\alpha \odot a) \oplus (\beta \odot a)$.

In the following let $a_0, a_1, b_0, b_1 \in \mathbb{R}$ and $\alpha \in \mathbb{R}$.

(a) Prove that H with the addition

$$\oplus : H \times H \rightarrow H : (a_0 + a_1 x, b_0 + b_1 x) \mapsto (a_0 + b_0) + (a_1 + b_1)x, \quad (34)$$

and scalar multiplication

$$\odot : \mathbb{R} \times H \rightarrow H : (\alpha, a_0 + a_1 x) \mapsto \alpha a_0 + \alpha a_1 x \quad (35)$$

is a vector space.

Solution: We start by proving that (H, \oplus) is a group, i.e., that there is a neutral element 0 with $f \oplus 0 = f$ for each $f \in H$. That there is an inverse f^{-1} with $f \oplus f^{-1} = 0$ for each $f \in H$ and that $(f \oplus g) \oplus h = f \oplus (g \oplus h)$. The neutral element is $0 = 0 + 0x$, i.e., for each $f = w_0 + w_1 x$ we have $f \oplus 0 = (w_0 + 0) + (w_1 + 0)x = w_0 + w_1 x$. For each $f = w_0 + w_1 x$ its inverse is $f^{-1} = -w_0 - w_1 x$, i.e., we have $f \oplus f^{-1} = w_0 - w_0 + (w_1 - w_1)x = 0$. Also for $f = a_0 + a_1 x, g = b_0 + b_1 x, h = c_0 + c_1 x$ we have

$$\begin{aligned} (f \oplus g) \oplus h &= a_0 + b_0 + (a_1 + b_1)x \oplus h \\ &= a_0 + b_0 + c_0 + (a_1 + b_1 + c_1)x \\ &= f \oplus b_0 + c_0 + (b_1 + c_1)x \\ &= f \oplus (g \oplus h), \end{aligned} \quad (36)$$

where the last two rows follow from the associativity of $+$. Thus, (H, \oplus) is a group.

Next, we go through the multiplicative axioms. Let $f = w_0 + w_1 x$. We have $0 \odot f = 0w_0 + 0w_1 x = 0$ and $1 \odot f = 1w_0 + 1w_1 x = f$. Let $\alpha, \beta \in \mathbb{R}$. Then,

$$\begin{aligned} (\alpha\beta) \odot f &= \alpha\beta w_0 + \alpha\beta w_1 x \\ &= \alpha(\beta w_0 + \beta w_1 x) \\ &= \alpha \odot (\beta \odot f). \end{aligned} \quad (37)$$

Finally, we go through the distributive axioms. Let $\alpha, \beta \in \mathbb{R}$ and $f = a_0 + a_1 x, g = b_0 + b_1 x \in H$. We have

$$\begin{aligned} \alpha \odot (f \oplus g) &= \alpha(a_0 + b_0) + \alpha(a_1 + b_1)x \\ &= \alpha a_0 + \alpha b_0 + (\alpha a_1 + \alpha b_1)x \\ &= (\alpha \odot f) \oplus (\alpha \odot g) \end{aligned} \quad (38)$$

and

$$\begin{aligned}
 (\alpha + \beta) \odot f &= (\alpha + \beta)a_0 + (\alpha + \beta)a_1x \\
 &= (\alpha a_0 + \beta a_0) + (\alpha a_1 + \beta a_1)x \\
 &= (\alpha \odot f) \oplus (\beta \odot g).
 \end{aligned} \tag{39}$$

Thus, H fulfills all vector space axioms and as a consequence is a vector space.

- (b) Prove the set $\{\phi_1, \phi_2\} \subset H$ with $\phi_1(x) = 1$, $\phi_2(x) = x$ is a basis of the linear vector space of affine functions (H, \oplus, \odot) .

Solution: We are going to prove by contradiction that ϕ_1 and ϕ_2 are linearly independent. Suppose they are not linearly independent. Then we have $c_1 1 + c_2 x = 0 \Leftrightarrow x = -\frac{c_1}{c_2}$, which is a contradiction since x is not a constant. Also, ϕ_1, ϕ_2 span the space H , as for each $f = w_0 + w_1 x \in H$ constants c_1, c_2 exist such that $f = c_1 \phi_1 + c_2 \phi_2$, namely, $c_1 = w_0$ and $c_2 = w_1$. Thus, $\{\phi_1, \phi_2\} \subset H$ is a basis of H .

- (c) Consider the hypothesis space of polynomials of degree k , i.e.,

$$H_k = \left\{ \sum_{i=0}^k w_i x^i : w_0, \dots, w_k \in \mathbb{R} \right\}. \tag{40}$$

Is H_k a linear vector space? If so, give a basis of H_k .

Solution: H_k is a vector space and $\{1, x, x^2, \dots, x^k\}$ is a basis of H_k .

We have seen that the space of affine functions is a two-dimensional linear vector space with basis functions $\phi_1(x) = 1$ and $\phi_2(x) = x$.

Similarly, we can define other p -dimensional vector spaces of functions simply by defining p basis functions $\phi_1 : \mathbb{R}^d \rightarrow \mathbb{R}, \dots, \phi_p : \mathbb{R}^d \rightarrow \mathbb{R}$. Each element f of the resulting hypothesis space H can be written as a linear combination of the chosen basis functions. That is, there exist real numbers $w_1, \dots, w_p \in \mathbb{R}$ such that $f(x) = \sum_{i=1}^p w_i \phi_i(x)$.