Exercice Sheet 3 – Regularization

Mathematics for Machine Learning

15 December 2025

Exercise 1. (Overparameterized regression in dim 1: linear versus Ridge regression.) Consider a regression problem in dimension 1, with only *one* train datapoint $(x_1, y_1) \in \mathbb{R} \times \mathbb{R}$. We take the linear model $\mathcal{F} = \{x \mapsto ax + b \mid (a, b) \in \mathbb{R}^2\}$.

1. **Linear regression (unregularized)**

   (a) On a plot, draw three different examples of empirical risk minimizers.

   (b) Prove that the set of ERMs for this problem is $\{(a, y_1 - ax_1) \mid a \in \mathbb{R}\}$, for any choice of cost $c \geq 0$ such that $c(y, y') = 0$ if and only if $y = y'$.

2. **Ridge regression**

   (a) What is the ERM for Ridge regression with parameter $\lambda > 0$, for the quadratic cost? *Hint: use the formula seen in the lecture. Recall the inversion formula for* 2×2 *matrices:* $\left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)^{-1} = \frac{1}{|ad-bc|}\left(\begin{smallmatrix} d & -c \\ -b & a \end{smallmatrix}\right)$.

   (b) Denote f_λ the obtained Ridge ERM. Is f_λ an ERM for the unregularized problem?

   (c) Check that f_0 (i.e. the limit of f_λ when $\lambda \to 0$) is an ERM for the unregularized problem, and that it has a minimal L2 norm $a^2 + b^2$ among such ERMs.

   (d) What is $\lim_{\lambda \to +\infty} f_\lambda$?

Exercise 2. (The effect of variable duplication) Consider a regression problem where we want to learn a label y from variables $x_1, \ldots, x_p$. For example, y can be the price of a painting, and x_1, x_2, x_3... its size, the average price of this type of paintings, its date of creation, etc.

Assume that the first variable is duplicated, i.e. that $x_1 = x_2$ in our data (both train and test). Then, for each painting, the features $(x_1, \ldots, x_p)$ take the form

$$\begin{pmatrix} \text{size} \\ \text{size} \\ \text{average price} \\ \vdots \end{pmatrix}.$$

The aim of this exercise is to investigate the effect of this variable duplication in linear regression and Ridge regression. Variable duplication is the simplest example of linear dependencies between variables.

1. **Linear regression**. Let $(a_1, \ldots, a_p) \in \mathbb{R}^p$ be an ERM for the quadratic cost, with train data $(x_1, y_1), \ldots, (x_n, y_n)$, i.e.

$$(a_1, \ldots, a_p) \in \operatorname{argmin}_{\alpha \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha_1 (x_i)_1 - \cdots - \alpha_p (x_i)_p)^2.$$

(a) Let $a'_1, a'_2 \in \mathbb{R}$ be such that $a'_1 + a'_2 = a_1 + a_2$. Prove that $(a'_1, a'_2, a_3, \ldots, a_p)$ is also an ERM.

(b) Deduce that there are infinitely many ERMs for the considered regression problem. Do these ERMs implement different predictions?

2. **Ridge regression.** We know from the lecture that there is a unique ERM for our regression problem when adding Ridge regularization. Let us denote it $(a_1, \ldots, a_p)$:

$$(a_1, \ldots, a_p) = \operatorname{argmin}_{\alpha \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha_1 (x_i)_1 - \cdots - \alpha_p (x_i)_p)^2 + \frac{\lambda}{2} \sum_{j=1}^{p} \alpha_j^2.$$

We want to prove that $a_1 = a_2$. Denote $L = \frac{1}{n} \sum_{i=1}^{n} (y_i - a_1 (x_i)_1 - \cdots - a_p (x_i)_p)^2$ the value of the unregularized empirical risk of $(a_1, \ldots, a_p)$.

(a) Let $a'_1, a'_2 \in \mathbb{R}$ be such that $a'_1 + a'_2 = a_1 + a_2$. Prove that the unregularized empirical risk of $(a'_1, a'_2, a_3, \ldots, a_p)$ is equal to L.

(b) Among all predictors in the set $\{(a'_1, a'_2, a_3, \ldots, a_p) \mid a'_1 + a'_2 = a_1 + a_2\}$, which one minimizes the regularization term $\frac{\lambda}{2}(a'^2_1 + a'^2_2 + \sum_{j=3}^{p} a_j^2)$? *Hint: denoting $C = a_1 + a_2$, we have $a'_2 = C - a'_1$. Then, the problem amounts to minimizing a degree 2 polynomial in a'_1.*

(c) Assuming that the set $\{(a'_1, a'_2, a_3, \ldots, a_p) \mid a'_1 + a'_2 = a_1 + a_2\}$ contains all predictors whose unregularized empirical risk is equal to L, deduce that $a_1 = a_2$. This illustrates the fact, stated in the lecture, that Ridge regression selects similar coefficients for similar variables.

Exercise 3. (Lasso for $X = I_p$.) Consider the lasso regression problem for $\lambda > 0$, in the special case where $n = p$ and $X = I_p$:

$$\operatorname{argmin} \frac{1}{2n} \|Y - A\|^2 + \lambda \|A\|_1. \tag{1}$$

The aim of this exercise is to solve this problem explicitly in this special case, i.e. find a formula for the argmin.

1. By rewriting (1) as a sum over $j = 1, \ldots, p$, prove that the p-dimensional minimization problem amounts to solving p one-dimensional problems of the form

$$\operatorname{argmin}_{a \in \mathbb{R}} \frac{1}{2n} (y - a)^2 + \lambda |a|. \tag{2}$$

2. Let us solve problem (2). Assume $y > 0$ and denote $\varphi \colon a \in \mathbb{R} \mapsto \frac{1}{2n}(y - a)^2 + \lambda |a|$.

   (a) Compute φ' on the intervals $(-\infty, 0)$ and $(0, +\infty)$.

   (b) Assume $0 < \lambda < y/n$. Solve (2).

   (c) Assume $\lambda = y/n$. Solve (2).

   (d) Assume $\lambda > y/n$. Solve (2).

3. If $y < 0$, prove that the solution of (2) is obtained by taking $\operatorname{argmin}_{a \in \mathbb{R}} \frac{1}{2n}(-y - a)^2 + \lambda |a|$ and changing its sign.

4. Deduce that the solution A^* of equation (1) is given by $A^* = S_{n\lambda}(Y)$, where $S_{n\lambda}$ is applied coordinate-wise, and is defined as

$$S_{n\lambda} \colon y \in \mathbb{R} \mapsto \operatorname{sgn}(y)\, \texttt{ReLU}(|y| - n\lambda).$$

Draw the graph of $S_{n\lambda}$. This function is called *soft thresholding*.

Solutions

Exercise 1.

1. (a) Seen in class.

   (b) *Note that the assumptions on the cost c have been corrected in this online version with respect to the printed version given in class.* We have seen in question 1.a) that there are empirical risk minimizers $f \in \mathcal{F}$ that exactly interpolate the data, i.e. such that $f(x_1) = y_1$. Such ERMs have a zero empirical risk. Therefore, the mimimal risk is 0, and $f \in \mathcal{F}$ is an ERM if and only if

   $$\hat{R}_n(f) = 0.$$

   By definition of the empirical risk, this can be rewritten equivalently as

   $$c(f(x_1), y_1) = 0,$$

   then as

   $$f(x_1) = y_1$$

   thanks to the assumption on the cost c. Denoting $f\colon x \mapsto ax + b$ by definition of the model $\mathcal{F}$, we obtain

   $$ax_1 + b = y_1,$$

   or equivalently:

   $$b = y_1 - ax_1.$$

   For each real number $a \in \mathbb{R}$, there is therefore one corresponding ERM $x \mapsto ax + y_1 - ax_1$, which proves the result.

2. Denoting $X = \begin{pmatrix} x_1 & 1 \end{pmatrix} \in \mathbb{R}^{1\times 2}$ and $Y = (y_1) \in \mathbb{R}^{1\times 1}$, we know from the lecture that the unique Ridge ERM $f_\lambda(x) =: a_\lambda x + b_\lambda$ is:

   $$\begin{pmatrix} a_\lambda \\ b_\lambda \end{pmatrix} = (X^\top X + \lambda I_2)^{-1} X^\top Y.$$

   Let us compute a_λ and b_λ explicitly. We have

   $$X^\top X + \lambda I_2 = \begin{pmatrix} x_1^2 + \lambda & x_1 \\ x_1 & 1 + \lambda \end{pmatrix}.$$

   Hence, with the matrix inversion formula:

   $$\begin{aligned}(X^\top X + \lambda I_2)^{-1} &= \frac{1}{|x_1^2 + \lambda(x_1^2 + 1) + \lambda^2 - x_1^2|} \begin{pmatrix} 1 + \lambda & -x_1 \\ -x_1 & x_1^2 + \lambda \end{pmatrix} \\ &= \frac{1}{\lambda(x_1^2 + 1 + \lambda)} \begin{pmatrix} 1 + \lambda & -x_1 \\ -x_1 & x_1^2 + \lambda \end{pmatrix}.\end{aligned}$$

   Moreover:

   $$X^\top Y = \begin{pmatrix} x_1 y_1 \\ y_1 \end{pmatrix}.$$

Going back to f_λ:

$$\begin{aligned}\begin{pmatrix} a_\lambda \\ b_\lambda \end{pmatrix} &= (X^\top X + \lambda I_2)^{-1} X^\top Y \\ &= \frac{1}{\lambda(x_1^2+1+\lambda)} \begin{pmatrix} 1+\lambda & -x_1 \\ -x_1 & x_1^2+\lambda \end{pmatrix} \begin{pmatrix} x_1 y_1 \\ y_1 \end{pmatrix} \\ &= \frac{1}{\lambda(x_1^2+1+\lambda)} \begin{pmatrix} \lambda x_1 y_1 \\ \lambda y_1 \end{pmatrix} \\ &= \frac{y_1}{x_1^2+1+\lambda} \begin{pmatrix} x_1 \\ 1 \end{pmatrix}.\end{aligned}$$

Therefore,

$$f_\lambda(x) = \frac{y_1}{x_1^2+1+\lambda}(x_1 x + 1).$$

3. We have

$$f_\lambda(x_1) = y_1 \frac{x_1^2+1}{x_1^2+1+\lambda}.$$

As $\lambda > 0$, the Ridge ERM f_λ is not an ERM for the unregularized problem.

4. With the previous answer, we check that $f_\lambda(x_1) \to y_1$ when $\lambda \to 0$, so f_0 is an ERM for the unregularized problem. Let us show that $(a_0, b_0) = \lim_{\lambda\to 0}(a_\lambda, b_\lambda)$ minimize the sum a^2+b^2 over the set of (unregularized) ERMs. Let $f(x) = ax + b$ be such an ERM. We have seen that $b = y_1 - ax_1$. Hence

$$a^2 + b^2 = a^2 + (y_1 - ax_1)^2 = (1+x_1^2)a^2 - 2x_1 y_1 a + y_1^2.$$

This a convex degree 2 polynomial in a. It is minimized for a that zeroes out the derivative, i.e.

$$a = \frac{x_1 y_1}{1+x_1^2},$$

which is precisely equal to $a_0 = \lim_{\lambda\to 0} a_\lambda$. This proves the result, as b_0 is determined by a_0.

5. When $\lambda \to +\infty$, we obtain $f_\lambda \to 0$. This is expected, as when $\lambda \to +\infty$, the regularization term takes over and the contribution of the empirical risk becomes negligible.

Exercise 2.

1. (a) We have

$$\frac{1}{n}\sum_{i=1}^n (y_i - a_1(x_i)_1 - a_2(x_i)_2 - \cdots - a_p(x_i)_p)^2 = \frac{1}{n}\sum_{i=1}^n (y_i - a_1'(x_i)_1 - a_2'(x_i)_2 - \cdots - a_p(x_i)_p)^2$$

as $a_1 + a_2 = a_1' + a_2'$. This means that the empirical risk is the same for $(a_1, \dots a_p)$ and $(a_1', a_2', a_3, \dots, a_p)$, hence both are empirical risk minimizers.

(b) For any $a_1' \in \mathbb{R}$, setting $a_2' = a_1 + a_2 - a_1'$ produces a new ERM, according to the previous question. There are therefore infinitely many ERMs for the considered regression problem. However, all these ERMs implement the same predictor, as the first and the second variable in the data are the same: for any $x \in \mathbb{R}^p$ such that $x_1 = x_2$, we have

$$\begin{aligned}a_1 x_1 + a_2 x_2 + \cdots + a_p x_p &= (a_1 + a_2)x_1 + a_3 x_3 + \cdots + a_p x_p \\ &= (a_1' + a_2')x_1 + a_3 x_3 + \cdots + a_p x_p \\ &= a_1' x_1 + a_2' x_2 + a_3 x_3 + \cdots + a_p x_p.\end{aligned}$$

2. (a) It is exactly the same argument as for question 1.a).

   (b) Denote $C = a_1 + a_2$ and $\mathcal{S} = \{(a'_1, a'_2, a_3, \dots, a_p) \mid a'_1 + a'_2 = a_1 + a_2\}$. Let $(a'_1, a'_2, a_3, \dots, a_p)$ be a predictor in the set $\mathcal{S}$. Then, this predictor can be rewritten as $(a'_1, C - a'_1, a_3, \dots, a_p)$. We want to minimize

$$a_1'^2 + (C - a'_1)^2 + \sum_{j=3}^{p} a_j^2$$

   over $a'_1 \in \mathbb{R}$. This is a convex degree 2 polynomial in a'_1. It is then minimized by a'_1 that zeroes out its derivative:

$$2a'_1 - 2(C - a'_1) = 0,$$

   which is equivalent to

$$a'_1 = \frac{C}{2}.$$

   The predictor in $\mathcal{S}$ that minimizes the regularization term is thus $(C/2, C/2, a_3, \dots, a_p)$.

   (c) The Ridge ERM $(a_1, \dots, a_p)$ has an empirical risk equal to L by definition, so it belongs to $\mathcal{S}$. Assume by contradiction that $a_1 \neq a_2$. Then, denoting $C = a_1 + a_2$, the predictor $(C/2, C/2, a_3, \dots, a_p)$ is also in $\mathcal{S}$: its empirical risk is equal to L. Moreover, its regularization term is strictly smaller than the one of $(a_1, \dots, a_p)$, according to the previous question, as it is the unique minimizer of the regularization term in $\mathcal{S}$. Finally

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - a_1(x_i)_1 - \cdots - a_p(x_i)_p)^2 + \frac{\lambda}{2}\sum_{j=1}^{p} a_j^2 >$$
$$\frac{1}{n}\sum_{i=1}^{n}(y_i - C/2(x_i)_1 - C/2(x_i)_2 - a_3x_3 - \cdots - a_p(x_i)_p)^2 + \frac{\lambda}{2}(\frac{C^2}{2} + \sum_{j=3}^{p} a_j^2),$$

   so we have found a predictor that has a smaller Ridge loss than the Ridge ERM, which leads to a contradiction.

Exercise 3.

1. We have

$$\frac{1}{2n}\|Y - A\|^2 + \lambda\|A\|_1 = \frac{1}{2n}\sum_{j=1}^{p}\left((y_j - a_j)^2 + \lambda|a_j|\right).$$

   To minimize the sum, we can minimize independently each term $(y_j - a_j)^2 + \lambda|a_j|$ by choosing the best $a_j \in \mathbb{R}$. We fall back on equation (2) with $a = a_j$ and $y = y_j$.

2. (a) If $a \in (-\infty, 0)$, then

$$\varphi'(a) = \frac{1}{n}(a - y) - \lambda.$$

   If $a \in (0, +\infty)$, then

$$\varphi'(a) = \frac{1}{n}(a - y) + \lambda.$$

   (b) If $0 < \lambda < y/n$, then φ' is < 0 on $(-\infty, 0)$ and, $\varphi'(a)$ changes sign from negative to positive on $(0, +\infty)$. Hence, φ reaches its minimum on $(0, +\infty)$, when φ' becomes 0, i.e. for

$$a^* = y - n\lambda > 0.$$

(c) When $y = \lambda/n$, φ' is still < 0 on $(-\infty, 0)$, and on $(0, +\infty)$, $\varphi' > 0$. So φ is decreasing ($=$ *strictement décroissante*) on $(-\infty, 0)$ and increasing ($=$ *strictement croissante*) on $(0, +\infty)$. As φ is continuous, it then reaches its minimum at $a^* = 0$.

(d) When, $\lambda > y/n$, φ' is still < 0 on $(-\infty, 0)$ as $y > 0$. Moreover, we still have $\varphi' > 0$ on $(0, +\infty)$. The conclusion is the same as in 2.c), i.e. $a^* = 0$.

3. If a^* minimizes $a \mapsto \frac{1}{2n}(-y - a)^2 + \lambda|a|$, then $-a^*$ minimizes

$$\alpha \mapsto \frac{1}{2n}(-y + \alpha)^2 + \lambda|\alpha|.$$

Equivalently,

$$-a^* = \text{argmin}_{\alpha \in \mathbb{R}} \frac{1}{2n}(y - \alpha)^2 + \lambda|\alpha|,$$

which proves the result.

4. We have seen that for each $j = 1, \ldots, p$,

$$\text{argmin}_{a_j \in \mathbb{R}} \frac{1}{2n}(y_j - a_j)^2 + \lambda|a_j| = \begin{cases} \text{sgn}(y_j)(|y_j| - n\lambda) & \text{if } \lambda < |y_j|/n \\ 0 & \text{if } \lambda > |y_j|/n \end{cases},$$

which is exactly the expected formula.

The graph of $S_{n\lambda}$ is as follows.