Math of ML : Exercises 2*

September 15, 2025

Unless otherwise specified, in the following exercises we consider a *fixed* design matrix $X \in \mathbb{R}^{n\times d}$ with rows $x_i^\top$ and assume that it is full-rank (in particular $n \geq d$).

Exercise 1 (Geometric interpretation of least-squares)**.** *Show that the vector of predictions $X\hat{w} = X(X^\top X)^{-1}X^\top y$ is the orthogonal projection of y on $\mathrm{im}(X) = \{Xw : w \in \mathbb{R}^d\}$.*

Solution 1. *Observe that $\ker(X^\top) = \{w \in \mathbb{R}^d : X^\top w = 0\}$ is the orthogonal complement of $\mathrm{im}(X)$. Thus, for any vector $y \in \mathbb{R}^n$ there exist $y_0 \in \mathrm{im}(X)$ and $y_\perp \in \ker(X^\top)$ such that*

$$y = y_0 + y_\perp.$$

The condition $y_0 \in \mathrm{im}(X)$ means, in particular, that for some $w_0 \in \mathbb{R}^d$ we have $y_0 = Xw_0$. It follows that

$$\begin{aligned} X(X^\top X)^{-1}X^\top y &= X(X^\top X)^{-1}X^\top(Xw_0 + y_\perp) \\ &= X(X^\top X)^{-1}X^\top Xw_0 + X(X^\top X)^{-1}X^\top y_\perp \\ &= Xw_0 + 0 \\ &= y_0, \end{aligned}$$

which is what we wanted to show.

Thus, we can interpret the OLS estimation as doing the following:

- *compute $\bar{y}$, the projection of y on the image of X;*
- *solve the linear system $Xw = \bar{y}$ which has a unique solution.*

Exercise 2 (Empirical risk of OLS)**.** *We consider the noisy measurement model $Y_i = x_i^\top w^* + Z_i$ where Z_i are independent, with zero mean $\mathbf{E}Z_i = 0$ and variance $\mathbf{E}[Z_i^2] = \sigma^2$ (the x_i are fixed). What is the expected empirical risk $\mathbf{E}[\hat{\mathcal{R}}_X(\hat{w})]$? Use this answer to propose an estimator of the noise variance σ^2 when $n > d$.*

Solution 2. *Let $Z \in \mathbb{R}^n$ denote the vector of noise variables such that the i-th entry is equal to Z_i. Let $y = Xw^* + Z$ and denote the "hat" matrix by $H = X(XX^\top)^{-1}X^\top$. Observing that $H^2 = H$ we have*

$$\begin{aligned} \mathbf{E}[\hat{\mathcal{R}}_X(\hat{w})] &= \frac{1}{n}\mathbf{E}\left[\|X\hat{w} - y\|_2^2\right] \\ &= \frac{1}{n}\mathbf{E}\left[\|(H - I)y\|_2^2\right] \\ &= \frac{1}{n}\mathbf{E}\left[y^\top(I - H)y\right] \\ &= \frac{1}{n}\mathbf{E}\left[\|y\|_2^2\right] - \frac{1}{n}\mathbf{E}[y^\top Hy] \end{aligned}$$

*Lénaïc Chizat EPFL lenaic.chizat@epfl.ch

We will now plug in the identity $y = Xw^* + Z$ and compute the resulting expectations. We have

$$\begin{aligned}
&\frac{1}{n}\mathbf{E}\left[\|y\|_2^2\right] - \frac{1}{n}\mathbf{E}[y^\top H y] \\
&= \frac{1}{n}\left((w^*)^\top X^\top X w^\star + \mathbf{E}[\|Z\|_2^2]\right) - \frac{1}{n}\left((w^*)^\top X^\top H X w^* + \mathbf{E}\left[Z^\top H Z\right]\right) \\
&= \frac{1}{n}\left((w^*)^\top X^\top X w^\star + n\sigma^2\right) - \frac{1}{n}\left((w^*)^\top X^\top X w^* + \mathbf{E}\left[Z^\top H Z\right]\right) \\
&= \sigma^2 - \frac{1}{n}\mathbf{E}\left[Z^\top H Z\right] \\
&= \sigma^2 - \frac{1}{n}\mathbf{E}\left[\operatorname{tr}\left(Z^\top H Z\right)\right] \\
&= \sigma^2 - \frac{1}{n}\mathbf{E}\left[\operatorname{tr}\left(H Z Z^\top\right)\right] \\
&= \sigma^2 - \frac{1}{n}\operatorname{tr}\left(H\sigma^2 I\right) \\
&= \sigma^2 - \frac{\sigma^2 d}{n}.
\end{aligned}$$

In particular, we have

$$\mathbf{E}[\hat{\mathcal{R}}_X(\hat{w})] = \frac{n-d}{n}\sigma^2.$$

Hence, whenever $n > d$, an unbiased *estimator of the noise variance σ^2 is given by $\frac{\|X\hat{w}-y\|_2^2}{n-d}$.*

Exercise 3 (OLS as a maximum-likelihood estimator)**.** *In this exercise, we make the stronger assumption that the i.i.d. noise random variables $Z_1, \ldots, Z_n$ all follow the Gaussian distribution $Z_i \sim \mathcal{N}(0, \sigma^2)$. Under the fixed-design Gaussian noise model, letting $Y = (Y_1, \ldots, Y_n)$, the likelihood that the observations Y were generated via the well-specified model $Y_i = x_i^\top w + Z_i$ is equal to*

$$L(Y|w, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-(Y_i - x_i^\top w)^2/(2\sigma^2)\Big).$$

The maximum likelihood estimator $(\tilde{w}, \tilde{\sigma})$ is defined as

$$(\tilde{w}, \tilde{\sigma}) = \operatorname{argmax}_{w\in\mathbb{R}^d, \sigma>0} L(Y|w, \sigma^2).$$

In this exercise, you are asked to:

1. *show that $\tilde{w}$ coincides with the OLS estimator;*
2. *compute the maximum likelihood estimator $\tilde{\sigma}^2$ for the variance. Is the estimator unbiased?*

Solution 3. *Maximizing the likelihood $L(Y|w, \sigma^2)$ over (w, σ^2) is the same as minimizing $-\log L(Y|w, \sigma^2)$ over (w, σ^2). Taking the negative logarithm and removing constants, the maximum likelihood estimator $(\tilde{w}, \tilde{\sigma}^2)$ becomes the minimizer of the objective*

$$\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - x_i^\top w)^2 + n\log(\sigma).$$

From the above expression, it is immediate that $\tilde{w}$ coincides with the OLS solution $\hat{w}$. Moreover, optimizing over σ^2, we have $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - x_i^\top \hat{w})^2$. Using the result of the previous exercise, observe that this estimator is biased towards 0*.*

Exercise 4 ((Practical exercise) Tuning ridge parameter)**.** *Recall that the ridge regression estimator is defined by*

$$\hat{w}_\lambda = \operatorname{argmin}_{w\in\mathbb{R}^d} \left\{ \frac{1}{n}\|Xw - y\|_2^2 + \lambda\|w\|_2^2 \right\}.$$

In Section 4 of Lecture 2, we suggested the choice of regularization parameter $\lambda^* = \frac{\sigma\sqrt{\operatorname{tr}\Sigma}}{\|w^*\|_2\sqrt{n}}$. *In particular, the suggested parameter* λ^* *scales as inverse square root of the number of samples.*

In the language of your choice, perform the following simulation for $n = 500, 600, 700, 800, 900, 1000$*:*

1. *Let* $d = 50$.
2. *Choose a fixed design matrix* $X \in \mathbb{R}^{n\times d}$ *by sampling each entry from a Gaussian distribution* $\mathcal{N}(0, 1)$.
3. *Let* $w^* = \mathbf{1}/\sqrt{d} \in \mathbb{R}^d$ *be the ground truth parameter of norm* $\|w^\star\|_2 = 1$.
4. *Fix some large enough grid of* λ *parameter values* $[\lambda_1, \ldots, \lambda_k]$.
5. *For each* $j = 1, \ldots, k$*, compute the expected excess risk of the ridge regression estimator with parameter* λ_j *using the bias-variance decomposition formula (see Lecture 2, Proposition 4.3; see also [Bach, 2024, Proposition 3.7])*

   For this part of the exercise, you may let the variance of the noise be equal to $\sigma^2 = 1$.
6. *Compute* λ_n^{opt} *that minimizes the excess risk among the values of* λ *in the grid* $[\lambda_1, \ldots, \lambda_k]$.

How does the computed solution λ_n^{opt} *compare with* λ^* *obtained theoretically in the lectures? As the number of samples* n *increases, how fast (approximately) does* λ_n^{opt} *decrease as a function of the number of samples* n*? Explain your findings.*

Solution 4. *First, observe that for* $n \gg d$ *the sampled fixed-design matrix* X *satisfies, with high probability,* $\frac{1}{n}X^\top X \approx I$. *Therefore, for the problem setup investigated in this exercise, the optimal regularization parameter* λ^* *suggested in the lectures satisfies*

$$\lambda^* \approx \sqrt{\frac{d}{n}}. \tag{1}$$

Moreover, the above choice of λ^* *yields the excess risk bound that decays as* $1/\sqrt{n}$ *as a function of the sample size* n *(such bounds are called "slow rate" bounds; they suggest that improving the excess risk twice requires quadrupling the amount of data).*

From the excess risk bounds obtained for the OLS estimator, which corresponds to $\lambda = 0$*, we already know that an excess risk bound of order* d/n *is achievable (such a bound is called "fast rate" bound; it suggests that improving the excess risk twice requires doubling the amount of data).*

Hence, the choice λ^* *given in* (1) overestimates *the amount of regularization that is required for the problem considered in this exercise. This is a consequence of the fact that the choice* λ^* *obtained in the lectures was computed based on* upper bounds *on the bias and variance terms, rather than their exact values (cf. [Bach, 2024, Proposition 3.7]).*

In the performed simulations, we find that the optimal choice λ_n^{opt} *decays approximately as* $1/n$ *as a function of the sample size.*

Exercise 5 (A non-linear estimator)**.** *In the previous exercises, we investigated the setting where the design matrix was fixed and the observations followed the model* $Y_i = x_i^\top w^* + Z_i$ *for some zero-mean noise variables* Z_i*.*

In this exercise, we take a look at the setting where the design is random and the Bayes optimal function is no longer assumed to be linear. Despite no longer assuming that the Bayes optimal function is linear, we may still be interested in bounding the excess risk

$$\mathbf{E}\mathcal{R}(\hat{f}) - \inf_{w \in \mathbb{R}^d} \mathcal{R}(\langle w, \cdot \rangle), \tag{2}$$

where recall that for any function $f : \mathbb{R}^d \to \mathbb{R}$ *we have*

$$\mathcal{R}(f) = \mathbf{E}_{(X,Y)\sim P}(f(X) - Y)^2.$$

In particular, the excess risk (2) *measures how close the estimator* $\hat{f}$ *gets to the best linear explanation of the data.*

The aim of this exercise is to introduce a new technique for bounding the excess risk, called average-stability *analysis, and further, to show how through the calculations performed in Exercise 6 it suggests a certain non-linear predictor due to Forster and Warmuth [2002].*

We shall make the following two assumptions[1] *on the unknown data generating mechanism:*

- *The data samples* (X_i, Y_i) *are generated i.i.d. from a distribution* P *such that we have* $|Y_i| \leq m$ *almost surely for some constant* $m > 0$*.*
- *For any data sample of size* n $(X_i, Y_i)_{i=1}^n$*, it holds with probability one that the matrix* $\sum_{i=1}^n X_i X_i^\top$ *is invertible (that is, we can compute the OLS estimator).*

This exercise is split into three parts.

1. *Denote a data sample of* $n+1$ *points by* $S_{n+1} = (X_i, Y_i)_{i=1}^{n+1}$ *and let* $S_{n+1}^{(-j)}$ *denote the sample* S_{n+1} *without the* j*-th input-output pair* (X_j, Y_j)*. Show that we may rewrite the expected risk of any estimator* $\hat{f} = \hat{f}[S_n]$ *as follows:*

$$\mathbf{E}_{S_n}\mathcal{R}(\hat{f}[S_n]) = \mathbf{E}_{S_{n+1}}(\hat{f}[S_n](X_{n+1}) - Y_{n+1})^2.$$

Show that for any estimator $\hat{f} = \hat{f}[(X_1, Y_1), \ldots, (X_n, Y_n)]$ *we have*

$$\mathbf{E}_{(X_i,Y_i)_{i=1}^n}\mathcal{R}(\hat{f}) = \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n}\left(\hat{f}[S_{n+1}^{(-j)}](X_j) - Y_j\right)^2\right].$$

2. *Let* $\hat{w} = \hat{w}[S_{n+1}]$ *be the OLS estimator computed on the sample* S_{n+1}*. Using the previous part of this exercise, show that the excess risk* (2) *can be upper bounded by*

$$\mathbf{E}\mathcal{R}(\hat{f}) - \inf_{w \in \mathbb{R}^d} \mathcal{R}(\langle w, \cdot \rangle) \leq \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n}(\hat{f}[S_{n+1}^{(-j)}](X_j) - Y_j)^2 - (X_j^\top \hat{w}[S_{n+1}] - Y_j)^2\right].$$

3. *Using part 2 of this exercise, suggest an estimator* $\hat{f}$ *such that*

$$\mathbf{E}\mathcal{R}(\hat{f}) - \inf_{w \in \mathbb{R}^d} \mathcal{R}(\langle \cdot, w \rangle) \leq \frac{2m^2 d}{n+1}.$$

[1] Both assumptions can be relaxed. The bounded response variable assumption can be weakened to assuming that the random variable $\mathbf{E}[Y^2|X]$ is almost surely bounded by m^2. We can get rid of the invertibility assumption completely, by replacing matrix inverses by their corresponding Moore-Penrose inverses.

You may use the fact that for any $j = 1, \dots, n+1$,

$$X_j^\top \hat{w}[S_{n+1}] = (1 - h_j) X_j^\top \hat{w}[S_{n+1}^{(-j)}] + h_j Y_j,$$

where $h_j \in [0,1]$ is the j-th leverage score[2] *defined by*

$$h_j = X_j^\top \left(\sum_{i=1}^{n+1} X_i X_i^\top \right)^{-1} X_j.$$

Solution 5.

1. *The first part of this exercise simply follows by relabelling the sample (X, Y) to (X_{n+1}, Y_{n+1}):*

$$\begin{aligned}
\mathbf{E}_{S_n} \mathcal{R}(\hat{f}(S_n)) &= \mathbf{E}_{S_n} \left[\mathbf{E}_{(X,Y)} \left[\left(\hat{f}(S_n)(X) - Y \right)^2 | S_n \right] \right] \\
&= \mathbf{E}_{S_n} \left[\mathbf{E}_{(X_{n+1},Y_{n+1})} \left[\left(\hat{f}(S_n)(X_{n+1}) - Y_{n+1} \right)^2 | S_n \right] \right] \\
&= \mathbf{E}_{S_{n+1}} \left(\hat{f}(S_n)(X_{n+1}) - Y_{n+1} \right)^2.
\end{aligned}$$

Using the fact that the sequence (X_i, Y_i) is i.i.d., we hence have, for any $j = 1, \dots, n+1$

$$\mathbf{E}_{S_{n+1}} \left(\hat{f}(S_n)(X_{n+1}) - Y_{n+1} \right)^2 = \mathbf{E}_{S_{n+1}} \left(\hat{f}(S_{n+1}^{(-j)})(X_j) - Y_j \right)^2.$$

Therefore, we have

$$\begin{aligned}
\mathbf{E}_{(X_i,Y_i)_{i=1}^n} \mathcal{R}(\hat{f}) &= \mathbf{E}_{S_{n+1}} \left(\hat{f}(S_n)(X_{n+1}) - Y_{n+1} \right)^2. \\
&= \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbf{E}_{S_{n+1}} \left(\hat{f}(S_n)(X_{n+1}) - Y_{n+1} \right)^2. \\
&= \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbf{E}_{S_{n+1}} \left(\hat{f}(S_{n+1}^{(-j)})(X_j) - Y_j \right)^2,
\end{aligned}$$

which is what we wanted to show.

2. *Using the result proved in the first part of this exercise and replacing the infimum over*

[2] https://en.wikipedia.org/wiki/Leverage_(statistics)

$w \in \mathbb{R}^d$ *via the OLS computed on the sample* S_{n+1} *yields*

$$\begin{aligned}
&\mathbf{E}\mathcal{R}(\hat{f}) - \inf_{w\in\mathbb{R}^d} \mathcal{R}(\langle w, \cdot\rangle) \\
&= \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\left(\hat{f}(S_{n+1}^{(-j)})(X_j) - Y_j\right)^2\right] - \inf_{w\in\mathbb{R}^d} \mathcal{R}(\langle w, \cdot\rangle) \\
&= \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\left(\hat{f}(S_{n+1}^{(-j)})(X_j) - Y_j\right)^2\right] - \inf_{w\in\mathbb{R}^d} \mathbf{E}_{(X,Y)}\left(X^\top w - Y\right)^2 \\
&= \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\left(\hat{f}(S_{n+1}^{(-j)})(X_j) - Y_j\right)^2\right] - \inf_{w\in\mathbb{R}^d} \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\left(X_j^\top w - Y_j\right)^2\right] \\
&\le \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\left(\hat{f}(S_{n+1}^{(-j)})(X_j) - Y_j\right)^2\right] - \mathbf{E}_{S_{n+1}}\left[\inf_{w\in\mathbb{R}^d}\frac{1}{n+1}\sum_{j=1}^{n+1}\left(X_j^\top w - Y_j\right)^2\right] \\
&= \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\left(\hat{f}(S_{n+1}^{(-j)})(X_j) - Y_j\right)^2 - \inf_{w\in\mathbb{R}^d}\frac{1}{n+1}\sum_{j=1}^{n+1}\left(X_j^\top w - Y_j\right)^2\right] \\
&= \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\left(\hat{f}(S_{n+1}^{(-j)})(X_j) - Y_j\right)^2 - \left(X_j^\top \hat{w} - Y_j\right)^2\right].
\end{aligned}$$

3. *Denoting* $\hat{w} = \hat{w}[S_{n+1}]$ *and* $\hat{w}_{(-j)} = \hat{w}[S_{n+1}^{(-j)}]$*, we have:*

$$X_j^\top \hat{w} - Y_j = (1-h_j)X_j^\top \hat{w}_{(-j)} - (1-h_j)Y_j.$$

Plugging the above expression into the second part of this exercise yields the following expected excess risk bound that is valid for any estimator $\hat{f}$*:*

$$\begin{aligned}
&\mathbf{E}\mathcal{R}(\hat{f}) - \inf_{w\in\mathbb{R}^d} \mathcal{R}(\langle w, \cdot\rangle) \\
&\le \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\left(\hat{f}(S_{n+1}^{(-j)})(X_j) - Y_j\right)^2 - (1-h_j)^2\left(X_j^\top \hat{w}_{(-j)} - Y_j\right)^2\right]. \qquad (3)
\end{aligned}$$

The above bound suggests the predictions of the form:

$$\hat{f}(S_{n+1}^{(-j)})(X_j) = (1-h_j)^2 X_j^\top \hat{w}_{(-j)}. \qquad (4)$$

The interpretation of the prediction rule (4) *(known in the literature as the Forster-Warmuth predictor) is that we* shrink *the predictions of OLS using the factor* $(1-h_j)^2$*, where* h_j *is the* statistical leverage *of the test point* (X_j, Y_j) *(note that computing the leverage* h_j does not *require knowing the value of response* Y_j*, which is the quantity that we are trying to predict). A point* X_j *has* high leverage*, intuitively, if it is not similar to the covariate vectors observed in training data sample. Thus, the Forster-Warmuth predictor makes more conservative predictions on "unfamiliar" regions of the covariate vectors. Finally, observe that computing Forster-Warmuth predictions can be done at a computational cost of* $\Theta(d^2)$ *operations (via the Sherman-Morrison formula),*

while computing the predictions of the OLS estimator can be done in $\Theta(d)$ number of operations.

To complete this exercise, we plug in the prediction rule (4) into the bound (3), yielding

$$\begin{aligned}
&\mathbf{E}\mathcal{R}(\hat{f}^{(\textit{Forster-Warmuth})}) - \inf_{w\in\mathbb{R}^d} \mathcal{R}(\langle w, \cdot\rangle) \\
&\leq \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\left((1-h_j)^2 X_j^\top \hat{w}_{(-j)} - Y_j\right)^2 - (1-h_j)^2\left(X_j^\top \hat{w}_{(-j)} - Y_j\right)^2\right] \\
&= \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\left((1-h_j)^2\left(X_j^\top \hat{w}_{(-j)} - Y_j\right) + (2h_j - h_j^2)(-Y_j)\right)^2 - (1-h_j)^2\left(X_j^\top \hat{w}_{(-j)} - Y_j\right)^2\right].
\end{aligned}$$

Using the facts that $h_j \in [0,1]$, that the quadratic function $x \mapsto x^2$ is convex, and that $|Y_j| \leq m$, we continue as follows:

$$\begin{aligned}
&= \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\left((1-h_j)^2\left(X_j^\top \hat{w}_{(-j)} - Y_j\right) + (2h_j - h_j^2)(-Y_j)\right)^2 - (1-h_j)^2\left(X_j^\top \hat{w}_{(-j)} - Y_j\right)^2\right] \\
&\leq \mathbf{E}_{S_{n+1}}\left[\frac{1}{n+1}\sum_{j=1}^{n+1}\left((1-h_j)^2 - (1-h_j)^2\right)\left(X_j^\top \hat{w}_{(-j)} - Y_j\right)^2 + (2h_j - h_j^2)Y_j^2\right] \\
&\leq \frac{2m^2}{n+1}\mathbf{E}_{S_{n+1}}\left[\sum_{j=1}^{n+1} h_j\right] \\
&= \frac{2m^2}{n+1}\mathbf{E}_{S_{n+1}}\left[\sum_{j=1}^{n+1} X_j^\top\left(\sum_{i=1}^{n+1} X_i X_i^\top\right)^{-1} X_j\right] \\
&= \frac{2m^2}{n+1}\mathbf{E}_{S_{n+1}}\left[\operatorname{tr}\left(\sum_{j=1}^{n+1} X_j^\top\left(\sum_{i=1}^{n+1} X_i X_i^\top\right)^{-1} X_j\right)\right] \\
&= \frac{2m^2}{n+1}\mathbf{E}_{S_{n+1}}\left[\operatorname{tr}\left(\left(\sum_{i=1}^{n+1} X_i X_i^\top\right)^{-1}\left(\sum_{j=1}^{n+1} X_j X_j^\top\right)\right)\right] \\
&\leq \frac{2m^2 d}{n+1}.
\end{aligned}$$

Exercise 6 (Optional: Stability of OLS predictions)**.** *Prove the fact used in part 3 of the previous exercise. That is to say:*

Let the fixed design matrix $X \in \mathbb{R}^{(n+1)\times d}$ contain $n+1$ rows $x_i^\top$ and let $y = (Y_1, \ldots, Y_{n+1}) \in \mathbb{R}^{n+1}$ denote the vector of observed response variables. Let $\hat{w} = (X^\top X)^{-1} X^\top y$ denote the OLS estimator computed on the $(n+1)$ data points.

Consider removing the data point (x_j, Y_j) and computing the OLS estimator

$$\hat{w}_{(-j)} = (X_{(-j)}^\top X_{(-j)})^\top X_{(-j)}^\top y_{(-j)},$$

where $X_{(-j)} \in \mathbb{R}^{n\times d}$ is a matrix obtained by removing the j-th row of X, and $y_{(-j)} \in \mathbb{R}^n$ is a vector obtained by removing the j-th entry of y.

Show that for any $j = 1, \dots, n+1$ it holds that

$$x_j^\top \hat{w} = (1 - h_j) x_j^\top \hat{w}_{(-j)} + h_j Y_j,$$

where $h_j \in [0, 1]$ is the j-th leverage score *defined by*

$$h_j = x_j^\top (X^\top X)^{-1} x_j.$$

Hint: *use the Sherman-Morrison*[3] *formula, stating that for any invertible square matrix $\Sigma \in \mathbb{R}^{d \times d}$ and any vector $x \in \mathbb{R}^d$ we have*

$$\left(\Sigma + xx^\top\right)^{-1} = \Sigma^{-1} - \frac{\Sigma^{-1} x x^\top \Sigma^{-1}}{1 + x^\top \Sigma^{-1} x}.$$

Solution 6. *Define*

$$\Sigma = \sum_{i=1}^{n+1} x_i x_i^\top, \quad \Sigma_{(-j)} = \Sigma - x_j x_j^\top$$

and

$$b = \sum_{i=1}^{n+1} x_i y_i, \quad b_{(-j)} = b - x_j y_j.$$

In particular, we have

$$\hat{w} = \Sigma^{-1} b \quad \text{and} \quad \hat{w}_{(-j)} = \Sigma_{(-j)}^{-1} b_{(-j)}.$$

By the Sherman-Morrison formula, it follows that

$$\begin{aligned}
x_j^\top \hat{w} &= x_j^\top \Sigma^{-1} b \\
&= x_j^\top \Sigma^{-1} b_{(-j)} + x_j^\top \Sigma^{-1} x_j y_j \\
&= x_j^\top \Sigma^{-1} b_{(-j)} + h_j y_j \\
&= x_j^\top \left(\Sigma_{(-j)} + x_j x_j^\top\right)^{-1} b_{(-j)} + h_j y_j \\
&= x_j^\top \left(\Sigma_{(-j)}^{-1} - \frac{\Sigma_{(-j)}^{-1} x_j x_j^\top \Sigma_{(-j)}^{-1}}{1 + x_j^\top \Sigma_{(-j)}^{-1} x_j}\right) b_{(-j)} + h_j y_j \\
&= x_j^\top \hat{w}_{(-j)} \left(1 - \frac{x_j^\top \Sigma_{(-j)}^{-1} x_j}{1 + x_j^\top \Sigma_{(-j)}^{-1} x_j}\right) + h_j y_j \\
&= x_j^\top \hat{w}_{(-j)} \left(\frac{1}{1 + x_j^\top \Sigma_{(-j)}^{-1} x_j}\right) + h_j y_j. \qquad (5)
\end{aligned}$$

Applying the Sherman-Morrison formula once again, we have

$$\begin{aligned}
h_j &= x_j^\top \Sigma^{-1} x_j \\
&= x_j^\top \left(\Sigma_{(-j)}^{-1} + x_j x_j^\top\right)^{-1} x_j \\
&= x_j^\top \Sigma_{(-j)}^{-1} x_j - \frac{\left(x_j^\top \Sigma_{(-j)}^{-1} x_j\right)^2}{1 + x_j^\top \Sigma_{(-j)}^{-1} x_j} \\
&= \frac{x_j^\top \Sigma_{(-j)}^{-1} x_j}{1 + x_j^\top \Sigma_{(-j)}^{-1} x_j} \\
&= 1 - \frac{1}{1 + x_j^\top \Sigma_{(-j)}^{-1} x_j}.
\end{aligned}$$

Plugging in the above identity into (5) yields the desired result.

[3] https://en.wikipedia.org/wiki/Sherman-Morrison_formula

References

Francis Bach. *Learning theory from first principles*. MIT press, 2024. URL https://www.di.ens.fr/~fbach/ltfp_book.pdf.

Jürgen Forster and Manfred K Warmuth. Relative expected instantaneous loss bounds. *Journal of Computer and System Sciences*, 64(1):76–102, 2002.