EPFL, Autumn 2024
Yoav Zemel
Guillaume Obozinski

Ho Yun
Shivang Sachar

Statistical Machine Learning

Exercise sheet 2

Exercise 2.1 (Continuation of Ex 1.1) Let $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ and $\mathbf{X}$ is a non-random full rank matrix of size $n \times p$. This setup contains the Gauss-Markov assumptions of a linear model.

(a) Prove the Gauss-Markov theorem, i.e, $\widehat{\boldsymbol{\beta}}$ is the best **linear unbiased** estimator (BLUE) of $\boldsymbol{\beta}$. "Best" in the sense that for all other linear unbiased estimators $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, $\mathrm{Cov}(\widetilde{\boldsymbol{\beta}}) - \mathrm{Cov}(\widehat{\boldsymbol{\beta}})$ is a positive semidefinite matrix.

Hints: Recall that an estimator $\widetilde{\boldsymbol{\beta}}$ is linear if $\widetilde{\boldsymbol{\beta}} = \mathbf{A}\boldsymbol{y}$, for some $\mathbf{A} \in \mathbb{R}^{p\times n}$. Notice that the matrix $\mathbf{A}$ can be decomposed as $\mathbf{A} = \mathbf{B} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.

Solution: There are many variants of the proof of the Gauss Markov theorem given in different textbooks, many of which rely on elementary results from linear algebra and some "tricks". We give here one such proof. First, note that the fact that we only consider estimators in the class of linear and unbiased estimators imposes the condition that $\mathbb{E}(\widetilde{\boldsymbol{\beta}}) = \mathbb{E}(\mathbf{A}\boldsymbol{y}) = \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \mathbb{E}(A\varepsilon) = \boldsymbol{\beta}$. This implies that $\mathbf{A}\mathbf{X} = \mathbf{I}$, the $p \times p$ identity matrix. Let $\mathbf{B} = \mathbf{A} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. We see that $\mathbf{B}\mathbf{X} = \mathbf{A}\mathbf{X} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X} = \mathbf{I} - \mathbf{I} = \mathbf{0}$, so

$$\mathrm{Cov}(\widetilde{\boldsymbol{\beta}}) = \mathrm{Cov}(\mathbf{A}\boldsymbol{y}) = \mathrm{Cov}(\mathbf{A}\boldsymbol{\varepsilon}) = \sigma^2\mathbf{A}\mathbf{A}^\top = \sigma^2\mathbf{B}\mathbf{B}^\top + \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1},$$

where the last equality follows from $\mathbf{B}\mathbf{X} = \mathbf{0}$. Using the result from (b), we see that $\mathrm{Cov}(\widetilde{\boldsymbol{\beta}}) - \mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2\mathbf{B}\mathbf{B}^\top$, where $\mathbf{B}\mathbf{B}^\top$ is now a positive semidefinite matrix. The matrix difference is the zero matrix if and only if $\mathbf{B} = \mathbf{0}$, that is, when $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. So indeed, we have that $\widehat{\boldsymbol{\beta}}$ gives the minimal variance in the class of all linear unbiased estimators.

(b) Assume now that the errors $\boldsymbol{\varepsilon}$ are normally distributed. Prove that $\widehat{\boldsymbol{\beta}}$ is the best estimator among **all unbiased** estimators. $\widehat{\boldsymbol{\beta}}$ is then a uniformly minimum variance unbiased (UMVU) estimator.

Hint: Remember the Cramér–Rao bound.

Solution: We have

$$\frac{\partial^2 \mathrm{RSS}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} = -2\mathbf{X}^\top\mathbf{X}.$$

The log-likelihood is $\ell(\boldsymbol{\beta}) = -\frac{1}{2\sigma^2}\mathrm{RSS} + \text{constant}$, hence the Fisher information matrix is

$$I(\widehat{\boldsymbol{\beta}}) = -\mathbb{E}\left(\frac{\partial^2\ell}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\right) = \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}.$$

Exercise 2.2 (The regression function) Recall that we are interested in the predictive model $f^* : \mathbb{R}^p \to \mathbb{R}$ that minimizes the expected error for the ℓ^2 loss. i.e., we want to find the function f^* such that

$$\mathbb{E}[\ell\{Y, f^*(\boldsymbol{X})\}] = \mathbb{E}[\{Y - f^*(\boldsymbol{X})\}^2] = \min_{f:\mathbb{R}^p\to\mathbb{R}} \mathbb{E}[\{Y - f(\boldsymbol{X})\}^2].$$

(a) Show that $f^*(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$.

Solution: There are many different ways to solve this exercise. We present here two solutions.

1st Solution: Let $m(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$. We see that

$$\begin{aligned}\mathbb{E}[L\{Y, f(\boldsymbol{X})\}] &= \mathbb{E}[\{Y - f(\boldsymbol{X})\}^2] \\ &= \mathbb{E}[\{Y - m(\boldsymbol{X}) + m(\boldsymbol{X}) - f(\boldsymbol{X})\}^2] \\ &= \mathbb{E}[\{Y - m(\boldsymbol{X})\}^2] + \mathbb{E}[\{m(\boldsymbol{X}) - f(\boldsymbol{X})\}^2] + 2\mathbb{E}[\{Y - m(\boldsymbol{X})\}\{m(\boldsymbol{X}) - f(\boldsymbol{X})\}] \\ &= \mathbb{E}[\{Y - m(\boldsymbol{X})\}^2] + \int_{\mathbb{R}^p} |m(\boldsymbol{x}) - f(\boldsymbol{x})|^2 P_{\boldsymbol{X}}(d\boldsymbol{x}),\end{aligned}$$

where $P_{\boldsymbol{X}}$ is the distribution of $\boldsymbol{X}$ and the last equality holds because

$$\begin{aligned}\mathbb{E}[\{Y - m(\boldsymbol{X})\}\{m(\boldsymbol{X}) - f(\boldsymbol{X})\}] &= \mathbb{E}(\mathbb{E}[\{Y - m(\boldsymbol{X})\}\{m(\boldsymbol{X}) - f(\boldsymbol{X})\}|\boldsymbol{X}]) \\ &= \mathbb{E}(\{m(\boldsymbol{X}) - f(\boldsymbol{X})\}\{\mathbb{E}(Y|\boldsymbol{X}) - m(\boldsymbol{X})\}) \\ &= 0.\end{aligned}$$

Clearly, the L^2 loss is minimized when $f(\boldsymbol{x}) = m(\boldsymbol{x})$, proving that indeed, $f^*(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$.

2nd Solution: Write

$$\mathbb{E}[\{Y - f(\boldsymbol{X})\}^2] = \mathbb{E}(\mathbb{E}[\{Y - f(\boldsymbol{X})\}^2 \mid \boldsymbol{X}]) = \int_{\mathbb{R}^p} \mathbb{E}[\{Y - f(\boldsymbol{X})\}^2 \mid \boldsymbol{X} = \boldsymbol{x}] P_{\boldsymbol{X}}(d\boldsymbol{x}),$$

with $\mathbb{E}[\{Y - f(\boldsymbol{X})\}^2 \mid \boldsymbol{X} = \boldsymbol{x}] \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^p$, thus it suffices to minimize $\mathbb{E}[\{Y - f(\boldsymbol{x})\}^2 \mid \boldsymbol{X} = \boldsymbol{x}]$ over $f(\boldsymbol{x})$ (almost everywhere, with respect to the measure $\mathbb{P}(\boldsymbol{X})$). The problem is equivalent to finding the value of $c = c(\boldsymbol{x}) \in \mathbb{R}$ that minimizes $\mathbb{E}\{(Y - c)^2 \mid \boldsymbol{X} = \boldsymbol{x}\}$ almost everywhere. Write

$$\begin{aligned}\mathbb{E}\{(Y - c)^2 \mid \boldsymbol{X} = \boldsymbol{x}\} &= \mathbb{E}(Y^2 - 2cY + c^2 \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \mathbb{E}(Y^2 \mid \boldsymbol{X} = \boldsymbol{x}) - 2c\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) + c^2.\end{aligned}$$

To minimize the previous expression, take the derivative with respect to c to find that the minimum is attained at $c = \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = m(\boldsymbol{x})$.

(b) If we consider the ℓ^1 loss instead, i.e., $\ell(y, \widehat{y}) = |y - \widehat{y}|$, what is f^*? (For simplicity suppose that $\mathbb{P}(Y \mid \boldsymbol{X})$ has a density.)

Solution: We want to find a function f^* that minimizes the expected loss $\mathbb{E}\{|Y - f(\boldsymbol{X})|\}$ over all functions $f : \mathbb{R}^p \to \mathbb{R}$. First, notice that

$$\mathbb{E}\{|Y - f(\boldsymbol{X})|\} = \mathbb{E}[\mathbb{E}\{|Y - f(\boldsymbol{X})| \mid \boldsymbol{X}\}]$$

Thus, using the same argument as in (a, 2nd solution), it suffices to find the constant c that minimizes $\mathbb{E}(|Y-c| \mid \boldsymbol{X}=\boldsymbol{x})$. Write

$$\begin{aligned}\mathbb{E}(|Y-c| \mid \boldsymbol{X}) &= \mathbb{E}\{(Y-c)I(Y-c\geq 0)+(c-Y)I(Y-c<0) \mid \boldsymbol{X}\}\\ &= \mathbb{E}\{YI(Y-c\geq 0) \mid \boldsymbol{X}\}-c\mathbb{P}(Y\geq c \mid \boldsymbol{X})\\ &\quad + c\mathbb{P}(Y<c \mid \boldsymbol{X})-\mathbb{E}\{YI(Y-c<0) \mid \boldsymbol{X}\}\\ &= \int_c^\infty yg(y \mid \boldsymbol{X})dy - c\mathbb{P}(Y\geq c \mid \boldsymbol{X})+c\mathbb{P}(Y<c \mid \boldsymbol{X})-\int_{-\infty}^c yg(y \mid \boldsymbol{X})dy,\end{aligned}$$

where g denotes the density of $Y \mid \boldsymbol{X}$. Thus

$$\begin{aligned}\frac{\partial}{\partial c}\mathbb{E}(|Y-c| \mid \boldsymbol{X}) &= -cg(c \mid \boldsymbol{X})-\mathbb{P}(Y\geq c \mid \boldsymbol{X})+cg(c \mid \boldsymbol{X})\\ &\quad +\mathbb{P}(Y<c \mid \boldsymbol{X})+cg(c \mid \boldsymbol{X})-cg(c \mid \boldsymbol{X})\\ &= -\mathbb{P}(Y\geq c \mid \boldsymbol{X})+\mathbb{P}(Y<c \mid \boldsymbol{X}),\end{aligned}$$

and the zero of that derivative occurs when $\mathbb{P}(Y<c \mid \boldsymbol{X})=\mathbb{P}(Y\geq c \mid \boldsymbol{X})=1/2$, i.e., for c equal to the conditional median $c=f^*(\boldsymbol{x})=\text{median}(Y|\boldsymbol{X}=\boldsymbol{x})$. Note: the same result holds without assuming a density for $Y \mid \boldsymbol{X}$, it is just more difficult to prove.

Exercise 2.3 (Bias-variance tradeoff) In this exercise, we consider the expected ℓ^2 error of a random predictive model $\widehat{f}_n$ (depends on a training set $\mathcal{D}_n$), defined as

$$\mathbb{E}\left[\int_{\mathbb{R}^p}\{\widehat{f}_n(\boldsymbol{x})-f^*(\boldsymbol{x})\}^2 P_{\boldsymbol{X}}(d\boldsymbol{x})\right]. \tag{1}$$

(a) For any random predictive model $\widehat{f}_n$ and any fixed point $\boldsymbol{x}_0\in\mathbb{R}^p$, prove that

$$\mathbb{E}\Big[\{\widehat{f}_n(\boldsymbol{x}_0)-f^*(\boldsymbol{x}_0)\}^2\Big]=\Big[\text{bias}\{\widehat{f}_n(\boldsymbol{x}_0)\}\Big]^2+\text{var}\{\widehat{f}_n(\boldsymbol{x}_0)\}.$$

Solution: We have that

$$\begin{aligned}\mathbb{E}\Big[\{\widehat{f}_n(\boldsymbol{x}_0)-f^*(\boldsymbol{x}_0)\}^2\Big] &= \mathbb{E}\Big[\{\widehat{f}_n(\boldsymbol{x}_0)-\mathbb{E}\{\widehat{f}_n(\boldsymbol{x}_0)\}+\mathbb{E}\{\widehat{f}_n(\boldsymbol{x}_0)\}-f^*(\boldsymbol{x}_0)\}^2\Big]\\ &= \Big[\text{bias}\{\widehat{f}_n(\boldsymbol{x}_0)\}\Big]^2+\text{var}\{\widehat{f}_n(\boldsymbol{x}_0)\}\end{aligned}$$

because

$$\mathbb{E}\Big[\{\widehat{f}_n(\boldsymbol{x}_0)-\mathbb{E}\{\widehat{f}_n(\boldsymbol{x}_0)\}\}\{\mathbb{E}\{\widehat{f}_n(\boldsymbol{x}_0)\}-f^*(\boldsymbol{x}_0)\}\Big]=0.$$

Note that in this expectation, $\mathbb{E}\{\widehat{f}_n(\boldsymbol{x}_0)\}-f^*(\boldsymbol{x}_0)$ is non-random. Important: $f^*(\boldsymbol{x}_0)$ is just a constant here, it could be the conditional mean, as in Exercise 2.1(a), or the conditional median, as in Exercise 2.1(b), or something else.

(b) Find a similar bias-variance decomposition for the expected ℓ^2 error (1).

Solution: By Fubini or Tonelli's theorem (where our function within the integrand is positive and measurable), we have that the expected ℓ^2 error is given by

$$\begin{aligned}\mathbb{E}\left[\int_{\mathbb{R}^p}\{\widehat{f}_n(\boldsymbol{x})-f^*(\boldsymbol{x})\}^2 P_{\boldsymbol{X}}(d\boldsymbol{x})\right] &= \int_{\mathbb{R}^p}\mathbb{E}\Big[\{\widehat{f}_n(\boldsymbol{x})-f^*(\boldsymbol{x})\}^2\Big]P_{\boldsymbol{X}}(d\boldsymbol{x})\\ &= \int_{\mathbb{R}^p}\Big[\text{bias}\{\widehat{f}_n(\boldsymbol{x})\}\Big]^2 P_{\boldsymbol{X}}(d\boldsymbol{x})+\int_{\mathbb{R}^p}\text{var}\{\widehat{f}_n(\boldsymbol{x})\}P_{\boldsymbol{X}}(d\boldsymbol{x}).\end{aligned}$$

We see that for the expected ℓ^2 error the squared bias and the variance integrate over the distribution of $\boldsymbol{X}$.

Exercise 2.4 (Ridge regression)

(a) Consider the linear regression model

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Define $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and the residuals as

$$r_i(\beta_0, \boldsymbol{\beta}) = y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, \ldots, n.$$

Show that the OLS estimator $\widehat{\beta}_0 = \overline{y} - \sum_{j=1}^{p} \beta_j x_{\cdot j}$ for any $\boldsymbol{\beta}$, where $x_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$. Hence deduce that

$$r_i(\widehat{\beta}_0, \boldsymbol{\beta}) = y_i - \overline{y} - \sum_{j=1}^{p} \beta_j (x_{ij} - x_{\cdot j}), \quad i = 1, \ldots, n.$$

Discuss the implications of this result.

Solution: The OLS estimator $\widehat{\beta}_0$ is found by minimizing the residual sum of squares with respect to β_0, keeping $\boldsymbol{\beta}$ fixed:

$$\text{RSS} = (\boldsymbol{y} - \beta_0 \mathbf{1} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \beta_0 \mathbf{1} - \mathbf{X}\boldsymbol{\beta}),$$

so $\frac{\partial}{\partial \beta_0}\text{RSS} = 2 \times \mathbf{1}^\top(\boldsymbol{y} - \beta_0 \mathbf{1} - \mathbf{X}\boldsymbol{\beta})$, and finding the root of the previous derivative gives $\widehat{\beta}_0 = \overline{y} - \sum_{j=1}^{p} \beta_j x_{\cdot j}$. The form of $r_i(\widehat{\beta}_0, \boldsymbol{\beta})$ is then trivial. Note that this result also follows from the Frisch–Waugh–Lovell (FWL) theorem. We now discuss the significance of this result. Note that in both residuals formulations, the vector $\boldsymbol{\beta}$ is the same. Hence, by centering the response and predictor variables it is always possible to get rid of the intercept β_0 in the first equation. The least squares estimation of $\boldsymbol{\beta}$ is the same in both the model with an intercept and in the model without it but with centered response and covariates. Thus $\boldsymbol{\beta}$ can first be estimated from the centered response and covariates (using OLS or another method) and β_0 will then be estimated by $\widehat{\beta}_0 = \overline{y} - \sum_{j=1}^{p} \widehat{\beta}_j x_{\cdot j}$. This motivates the definition of the ridge estimator in the next equation (equation (2)), compared to the definition with an unconstrained intercept: as β_0 in unconstrained we know $\widehat{\beta}_0 = \overline{y} - \sum_{j=1}^{p} \widehat{\beta}_j x_{\cdot j}$ and $\widehat{\boldsymbol{\beta}}$ is found solving equation (2), thus the two formulations of ridge will give the same estimates of $\boldsymbol{\beta}$ and β_0.

(b) Define the ridge regression estimator as a minimizer of the penalized residual sum of squares,

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}, \tag{2}$$

where $\lambda \geq 0$ is a parameter that controls the amount of shrinkage. Show that the ridge regression solution always exists, even if $\mathbf{X}$ does not have full rank, and is given by

$$\widehat{\boldsymbol{\beta}}(\lambda) = (\mathbf{X}^\top \mathbf{X} + n\lambda \mathbf{I})^{-1} \mathbf{X}^\top \boldsymbol{y}.$$

Note that the ridge estimator is still linearly depending on the response $\boldsymbol{y}$, as for ordinary least squares.

Solution: Taking gradient of $\frac{1}{n}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}$ with respect to $\boldsymbol{\beta}$, setting it to 0, and rearranging the terms, gives the normal equations

$$(\mathbf{X}^\top \mathbf{X} + n\lambda \mathbf{I})\widehat{\boldsymbol{\beta}}(\lambda) = \mathbf{X}^\top \boldsymbol{y}.$$

Now note that the matrix $\mathbf{X}^\top \mathbf{X} + n\lambda \mathbf{I}$ is always invertible: $\mathbf{X}^\top \mathbf{X}$ is positive semidefinite (since $\boldsymbol{x}^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{x} = \|\mathbf{X}\boldsymbol{x}\| \geq 0$) hence its eigenvalues $\alpha_i \geq 0$, and the eigenvalues of $\mathbf{X}^\top \mathbf{X} + n\lambda \mathbf{I}$ are $\alpha_i + n\lambda > 0$ for any $\lambda > 0$. Hence the ridge estimator always exists and is uniquely defined as

$$\widehat{\boldsymbol{\beta}}(\lambda) = (\mathbf{X}^\top \mathbf{X} + n\lambda \mathbf{I})^{-1} \mathbf{X}^\top \boldsymbol{y}.$$

(c) Show that the ridge regression estimator defined in (2) equals

$$\widehat{\boldsymbol{\beta}}(t) = \underset{\|\boldsymbol{\beta}\|^2 \leq t}{\operatorname{argmin}} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \tag{3}$$

for a given $t = t(\lambda)$. *Hint: Use the Karush–Kuhn–Tucker (KKT) method.*

Solution: The constrained optimization problem in (3) can be solved by means of the Karush–Kuhn–Tucker (KKT) multiplier method, which minimizes a function subject to inequality constraints. The KKT multiplier method states that, under some regularity conditions (here satisfied), there exists a unique λ, called the multiplier, such that the solution $\widehat{\boldsymbol{\beta}}$ of the constrained minimization problem (3) satisfies the so-called KKT conditions. Define the Lagrangian of the problem as

$$\ell(\boldsymbol{\beta}, \lambda) = \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \left(\|\boldsymbol{\beta}\|^2 - t \right).$$

The KKT conditions are:

1. $\dfrac{\partial \ell}{\partial \boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}, \lambda) = \mathbf{0}$.
2. $\lambda \geq 0$.
3. $\lambda(\|\widehat{\boldsymbol{\beta}}\|^2 - t) = 0$.
4. $\|\widehat{\boldsymbol{\beta}}\|^2 - t \leq 0$.

Now, suppose that the solution that minimizes the minimization problem in (2) is given by the ridge estimate $\widehat{\boldsymbol{\beta}}(\lambda/n)$. For $t = \|\widehat{\boldsymbol{\beta}}(\lambda/n)\|^2$ it is clear that $\widehat{\boldsymbol{\beta}}(\lambda/n)$ satisfies all KKT conditions. Hence, both the constrained optimization problem in (3) and the minimization problem in (2) have the same solution when $t = \|\widehat{\boldsymbol{\beta}}(\lambda/n)\|^2$.

Exercise 2.5 The Gauss-Markov Theorem makes the assumption that the training data is generated as $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathbf{X}$ is a non-random full rank matrix of size $n \times p$, where $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$.

(a) Explain why the Gauss-Markov Theorem still holds for any random design matrix $\mathbf{X}$ (in particular without assuming that the rows of $\mathbf{X}$ are i.i.d.) provided we change the assumptions and assume that $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{I}$.

Solution: Recall that when the design matrix $\mathbf{X}$ is non-random, the Gauss-Markov theorem indicates that for any unbiased linear estimator $\tilde{\beta}$, there exists $\boldsymbol{K}_{\mathbf{X}}$ positive semi-definite such that

$$\mathrm{Cov}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \mathrm{Cov}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) + \boldsymbol{K}_{\mathbf{X}}, \qquad (\star)$$

where the covariances are conditional covariances and $\boldsymbol{K}_{\mathbf{X}}$ does not only depend on the choice of the estimator $\tilde{\boldsymbol{\beta}}$ but also on $\mathbf{X}$. To obtain a statement which is again expressed in terms of the marginal covariance of the estimators, we can use the variance decomposition formula

$$\mathrm{Cov}(\tilde{\boldsymbol{\beta}}) = \mathbb{E}[\mathrm{Cov}(\tilde{\boldsymbol{\beta}}|\mathbf{X})] + \mathrm{Cov}(\mathbb{E}[\tilde{\boldsymbol{\beta}}|\mathbf{X}]),$$

so taking expectation in $(\star)$, we get

$$\begin{aligned}\mathrm{Cov}(\tilde{\boldsymbol{\beta}}) &= \mathbb{E}[\mathrm{Cov}(\widehat{\boldsymbol{\beta}}|\mathbf{X})] + \mathbb{E}[\boldsymbol{K}_{\mathbf{X}}] + \mathrm{Cov}(\mathbb{E}[\tilde{\boldsymbol{\beta}}|\mathbf{X}]) \\ &= \mathrm{Cov}(\widehat{\boldsymbol{\beta}}) + \mathbb{E}[\boldsymbol{K}_{\mathbf{X}}] + \mathrm{Cov}(\mathbb{E}[\tilde{\boldsymbol{\beta}}|\mathbf{X}]),\end{aligned}$$

where we use the fact that $\mathbb{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] = \beta$ thus $\mathrm{Cov}(\mathbb{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]) = 0$. Finally, both $\mathbb{E}[\boldsymbol{K}_{\mathbf{X}}]$ and $\mathrm{Cov}(\mathbb{E}[\tilde{\boldsymbol{\beta}}|\mathbf{X}])$ are also positive semi-definite, which yields the Gauss-Markov theorem.

(b) Let $\tilde{\boldsymbol{\beta}}$ be any linear unbiased estimator and let $\widehat{\boldsymbol{\beta}}$ be the linear regression estimator (aka ordinary least squares estimator). Show that as a consequence of the Gauss-Markov theorem:

$$\forall \boldsymbol{x} \in \mathbb{R}^p, \qquad \mathrm{Var}(\boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}) \leq \mathrm{Var}(\boldsymbol{x}^\top\tilde{\boldsymbol{\beta}}).$$

Solution: By the previous question, we have that

$$\mathrm{Cov}(\tilde{\boldsymbol{\beta}}) = \mathrm{Cov}(\widehat{\boldsymbol{\beta}}) + \boldsymbol{K},$$

with $\boldsymbol{K} = \mathbb{E}[\boldsymbol{K}_{\mathbf{X}}]$. But this entails that, for any fixed $\boldsymbol{x} \in \mathbb{R}^p$, we have

$$\boldsymbol{x}^\top\mathrm{Cov}(\tilde{\boldsymbol{\beta}})\boldsymbol{x} = \boldsymbol{x}^\top\mathrm{Cov}(\widehat{\boldsymbol{\beta}})\boldsymbol{x} + \boldsymbol{x}^\top\boldsymbol{K}\boldsymbol{x};$$

now since $\boldsymbol{K}$ is positive semi-definite, we must have $\boldsymbol{x}^\top\boldsymbol{K}\boldsymbol{x} \geq 0$ and it is easy to check that $\mathrm{Var}(\boldsymbol{x}^\top\tilde{\boldsymbol{\beta}}) = \boldsymbol{x}^\top\mathrm{Cov}(\tilde{\boldsymbol{\beta}})\boldsymbol{x}$ using the definitions of variance and covariance. Hence the result.

(c) Consider now i.i.d. data (X_i, Y_i) with $Y_i = X_i^\top\boldsymbol{\beta} + \varepsilon_i$, $\mathbb{E}[\varepsilon_i|X_i] = 0$ and $\mathrm{Var}(\varepsilon_i|X_i) = \sigma^2$. For data following this distribution, express the target function for the quadratic risk as a function of $\boldsymbol{\beta}$.

Solution: We know that for the quadratic risk $f^*(x) = \mathbb{E}[Y|X = x]$ and

$$f^*(X_i) = \mathbb{E}[Y_i|X_i] = \mathbb{E}[X_i^\top \boldsymbol{\beta} + \varepsilon_i|X_i] = X_i^\top \boldsymbol{\beta} + \mathbb{E}[\varepsilon_i|X_i] = X_i^\top \boldsymbol{\beta}.$$

So $f^*(\boldsymbol{x}) = \boldsymbol{\beta}^\top \boldsymbol{x}$.

(d) Let $\widehat{f} : \boldsymbol{x} \mapsto \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}$ and $\tilde{f} : \boldsymbol{x} \mapsto \boldsymbol{x}^\top \tilde{\boldsymbol{\beta}}$ for $\tilde{\boldsymbol{\beta}}$ some unbiased linear estimator based on $\mathbf{X}$ and $\boldsymbol{y}$. Show that for any such $\tilde{f}$, if $\mathcal{R}$ denotes the quadratic risk (i.e. the risk associated with the square loss), then we necessarily have $\mathbb{E}[\mathcal{R}(\widehat{f})] \leq \mathbb{E}[\mathcal{R}(\tilde{f})]$. Show that the same inequality actually holds conditionally on the value of any new $X = \boldsymbol{x}$.

Solution: Note first that since $\tilde{\boldsymbol{\beta}}$ is unbiased then, for any fixed $\boldsymbol{x}$.

$$\mathbb{E}[\tilde{f}(\boldsymbol{x})] = \mathbb{E}[\tilde{\boldsymbol{\beta}}^\top \boldsymbol{x}] = \mathbb{E}[\tilde{\boldsymbol{\beta}}]^\top \boldsymbol{x} = \boldsymbol{\beta}^\top \boldsymbol{x} = f^*(\boldsymbol{x}),$$

and, since $\widehat{\boldsymbol{\beta}}$ is itself unbiased, $\mathbb{E}[\widehat{f}(\boldsymbol{x})] = f^*(\boldsymbol{x})$. Now, this entails that

$$\mathrm{Var}(\tilde{f}(\boldsymbol{x})) = \mathbb{E}[(\tilde{f}(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2].$$

To connect this with the risk, we denote by $D_n = \{(X_i, Y_i)\}_{i=1..n}$ the training set and we use the result established on the quadratic risk in class:

$$\mathbb{E}[(\tilde{f}(\boldsymbol{x}) - Y)^2|X = \boldsymbol{x}, D_n] = (\tilde{f}(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2 + \mathbb{E}[(f^*(\boldsymbol{x}) - Y)^2|X = \boldsymbol{x}, D_n]$$

that we can also write

$$\mathcal{R}(\tilde{f}(\boldsymbol{x})|\boldsymbol{x}) = (\tilde{f}(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2 + \mathcal{R}(f^*(\boldsymbol{x})|\boldsymbol{x}),$$

by identifying the conditional risks of $\tilde{f}$ and f^*. Now taking expectations w.r.t. to D_n on both sides we get

$$\mathbb{E}[\mathcal{R}(\tilde{f}(\boldsymbol{x})|\boldsymbol{x})] = \mathrm{Var}(\tilde{f}(\boldsymbol{x})) + \mathcal{R}(f^*(\boldsymbol{x})|\boldsymbol{x}),$$

where $\mathbb{E}[\mathcal{R}(\tilde{f}(\boldsymbol{x})|\boldsymbol{x})]$ is the expected conditional risk at $\boldsymbol{x}$, with an expectation taken over the choice of the training data D_n only. Given that we have the same formula hold for $\widehat{f}$ and given that $\mathrm{Var}(\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}) \leq \mathrm{Var}(\boldsymbol{x}^\top \tilde{\boldsymbol{\beta}})$ we have established that

$$\mathbb{E}[\mathcal{R}(\widehat{f}(\boldsymbol{x})|\boldsymbol{x})] \leq \mathbb{E}[\mathcal{R}(\tilde{f}(\boldsymbol{x})|\boldsymbol{x})].$$

By taking expectations over the choice of $X = \boldsymbol{x}$, we get

$$\mathbb{E}[\mathcal{R}(\widehat{f})] \leq \mathbb{E}[\mathcal{R}(\tilde{f})].$$