In the following questions, where appropriate, suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. and take values in $\mathcal{X} \times \mathcal{Y}$. We will take $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \{-1, 1\}$ and the loss ℓ will be misclassification loss, unless it is specified that a regression setting is being considered, in which case the loss will typically be squared error. Assume that the computational complexity of inverting $M \in \mathbb{R}^{m\times m}$ is $O(m^3)$, and forming BC where $B \in \mathbb{R}^{a\times b}$ and $C \in \mathbb{R}^{b\times c}$ is $O(abc)$.

1. Show that

$$R(h) - R(h_0) = \mathbb{E}\{\mathbb{1}_{\{h(X)\neq h_0(X)\}}|2\eta(X) - 1|\}$$

where

$$h_0(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ -1 & \text{otherwise} \end{cases}$$

and $\eta(x) := \mathbb{P}(Y = 1 \,|\, X = x)$.
Solution: We have

$$\mathbb{P}(Y \neq h(X)|X = x) = \mathbb{1}_{\{h(x)=-1\}}\eta(x) + \mathbb{1}_{\{h(x)=1\}}(1 - \eta(x)),$$

so,

$$\begin{aligned} \mathbb{P}(Y \neq h(X)|X = x) - \mathbb{P}(Y \neq h_0(X)|X = x) &= \mathbb{1}_{\{h(x)=1,h_0(x)=-1\}}(1 - 2\eta(x)) + \mathbb{1}_{\{h(x)=-1,h_0(x)=1\}}(2\eta(x) - 1) \\ &= \mathbb{1}_{\{h(x)\neq h_0(x)\}}|2\eta(x) - 1| \end{aligned}$$

using the definition of h_0 for the final equality. Taking expectations we then obtain the desired result.

2. In each of the settings below, find a classifier that minimises the risk corresponding to the loss functions given.

(a) Consider the weighted misclassification loss $\ell : \{-1, 1\}^2 \to \mathbb{R}$ given by $\ell(-1, -1) = \ell(1, 1) = 0$ and $\ell(-1, 1) = \alpha$, $\ell(1, -1) = \beta$ where $\alpha, \beta > 0$.
Solution: Using the argument from the previous question, we have for any classifier h that

$$\mathbb{E}(\ell(h(X), Y)|X = x) = \alpha\mathbb{1}_{\{h(x)=-1\}}\eta(x) + \beta\mathbb{1}_{\{h(x)=1\}}(1 - \eta(x)).$$

To minimise the risk it suffices to pick h such that $h(x)$ minimises the RHS of the above i.e. we may take

$$h(x) = \begin{cases} 1 & \text{if } \beta(1 - \eta(x)) < \alpha\eta(x), \ \ (\text{ so } \eta(x) > \beta/(\alpha + \beta)\ ) \\ -1 & \text{otherwise.} \end{cases}$$

(b) Suppose $\mathcal{Y} = \{1, \ldots, K\}$ and loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfies

$$\ell(y', y) = \begin{cases} 0 & \text{if } y = y' \\ 1 & \text{otherwise.} \end{cases}$$

Solution: We have

$$\mathbb{E}(\ell(h(X), Y)|X = x) = \sum_{k=1}^{K} \mathbb{P}(Y = k|X = x)(1-\mathbb{1}_{\{h(x)=k\}}) = 1-\sum_{k=1}^{K} \mathbb{P}(Y = k|X = x)\mathbb{1}_{\{h(x)=k\}}.$$

To minimise the risk, it suffices to pick h such that $h(x)$ minimises the RHS of the above, so we should take

$$h(x) \in \operatorname*{argmax}_k \mathbb{P}(Y = k|X = x).$$

3. Let $\hat{h} = \hat{h}_D$ be a hypothesis trained on data $D = (X_i, Y_i)_{i=1}^n$ formed of iid copies of an independent random pair (X, Y). Define $\tilde{h}_{X_{1:n}}(x) := \mathbb{E}(\hat{h}_D(x) \mid X_{1:n})$.

   (a) Show that

$$\mathbb{E}[\{Y - \hat{h}_D(X)\}^2 \mid X = x] = \mathbb{E}\{\mathbb{E}(Y \mid X = x) - \tilde{h}_{X_{1:n}}(x)\}^2 + \mathbb{E}\{\hat{h}_D(x) - \tilde{h}_{X_{1:n}}(x)\}^2 + \mathrm{Var}(Y \mid X = x).$$

   Solution: We have (from lectures),

$$\mathbb{E}[\{Y - \hat{h}_D(X)\}^2 \mid X] = \mathbb{E}[\{\mathbb{E}(Y \mid X) - \hat{h}_D(X)\}^2 \mid X] + \mathrm{Var}(Y \mid X).$$

   Next, using

$$\mathbb{E}\{Z - f(W)\}^2 = \mathbb{E}\{Z - \mathbb{E}(Z \mid W)\}^2 + \mathbb{E}\{\mathbb{E}(Z \mid W) - f(W)\}^2.$$

   with $W = (X, X_{1:n})$, $f(W) = \mathbb{E}(Y \mid X)$ and $Z = \hat{h}_D(X)$, we have

$$\mathbb{E}[\{\mathbb{E}(Y \mid X) - \hat{h}_D(X)\}^2 \mid X] = \mathbb{E}[\{\mathbb{E}(Y \mid X) - \tilde{h}_{X_{1:n}}(X)\}^2 \mid X] + \mathbb{E}[\{\hat{h}_D(X) - \tilde{h}_{X_{1:n}}(X)\}^2 \mid X].$$

   Then 'fixing what is known' gives the result.

   (b) Show that considering squared error loss,

$$\mathbb{E}R(\hat{h}_D) - \mathbb{E}R(\tilde{h}_{X_{1:n}}) = \mathbb{E}\{\hat{h}_D(X) - \tilde{h}_{X_{1:n}}(X)\}^2.$$

   Solution: Follows easily from considering a decomposition as in (a) but with $\hat{h}_D = \tilde{h}_{X_{1:n}}$ (there is no restriction on $\hat{h}_D$, so this is permitted).

4. Consider performing OLS regression using a set of d basis functions $(\varphi_1, \ldots, \varphi_d) := \varphi$ using data $(X_i, Y_i)_{i=1}^n$. Assume that the matrix $\Phi \in \mathbb{R}^{n \times d}$ with ith row $\varphi(X_i) \in \mathbb{R}^d$ has full column rank.

   (a) Show that the OLS coefficient vector $\hat{\beta} \in \mathbb{R}^d$ may be obtained in $O(nd^2)$ operations.
   Solution: Computing $\Phi^\top \Phi$ is $O(nd^2)$ and inverting this is $O(d^3)$ (but note $d \leq n$ as Φ has full column rank). Next computing $\Phi^\top Y_{1:n} \in \mathbb{R}^d$ is $O(nd)$ and then $(\Phi^\top \Phi)^{-1} (\Phi^\top Y_{1:n})$ is $O(d^2)$. Thus the overall complexity is $O(nd^2)$.

   (b) Show that the leave-one-out cross-validation score

$$\frac{1}{n}\sum_{i=1}^n \{Y_i - \varphi(X_i)^\top \hat{\beta}_{-i}\}^2$$

   may be computed in $O(nd^2)$ operations. Here $\hat{\beta}_{-i} \in \mathbb{R}^d$ is the OLS coefficient vector when performing regression using a dataset with the ith point removed. [Use the matrix identity

$$(A - bb^\top)^{-1} = A^{-1} + \frac{A^{-1}bb^\top A^{-1}}{1 - b^\top A^{-1} b}$$

   whenever $A \in \mathbb{R}^{p \times p}$ is invertible, $b \in \mathbb{R}^p$ and $b^\top A^{-1} b \neq 1$. Also assume $\varphi(X_i)^\top (\Phi^\top \Phi)^{-1} \varphi(X_i) < 1$, which holds provided each $(n-1) \times d$ sub-matrix of Φ has full column rank.] [*Hint: Consider first computing* $(\Phi^\top \Phi)^{-1} \varphi(X_i) \in \mathbb{R}^d$ *for all* $i = 1, \ldots, n$.]
   Solution: Computing $a_i := (\Phi^\top \Phi)^{-1} \varphi(X_i)$ for all i is $O(d^3 + nd^2) = O(nd^2)$. Now, writing $A := \Phi^\top \Phi$, $\phi_i := \varphi(X_i)$, $y := Y_{1:n}$

$$\begin{aligned} \phi_i^\top \hat{\beta}_{-i} &= \phi_i^\top (A - \phi_i \phi_i^\top)^{-1} (\Phi^\top y - \phi_i Y_i) \\ &= \left(a_i^\top + \frac{a_i^\top \phi_i a_i^\top}{1 - \phi_i^\top a_i} \right) (\Phi^\top y - \phi_i Y_i). \end{aligned}$$

   Now $a_i^\top \phi_i$ and $a_i^\top \Phi^\top y$ are both $O(d)$ computations, provided $\Phi^\top y$ has already been computed ($O(nd)$ time). Thus in total, computing all $\varphi(X_i)^\top \hat{\beta}_{-i}$ costs $O(nd^2)$ and hence the CV score above may be computed in $O(nd^2)$ time.

5. Consider a regression setting as in the previous question with $\Phi \in \mathbb{R}^{n\times d}$ and φ defined as above. For $\lambda \geq 0$, consider $\hat{h}_\lambda$ given by $\hat{h}_\lambda(x) = \varphi(x)^\top \hat{\beta}_\lambda$ with

$$\hat{\beta}_\lambda := \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \{\|Y_{1:n} - \Phi\beta\|_2^2 + \lambda\|\beta\|_2^2\}.$$

(a) Show that $\hat{\beta}_\lambda = (\Phi^\top\Phi + \lambda I)^{-1}\Phi^\top Y_{1:n}$.
Solution: We can differentiate the objective w.r.t. β to obtain

$$\Phi^\top(Y_{1:n} - \Phi\hat{\beta}_\lambda) = \lambda\hat{\beta}_\lambda,$$

which easily yields the result.

(b) Suppose $\mathrm{Var}(Y_1 \,|\, X_1 = x) > 0$ is constant in x and $\varphi(x)$ is not the zero vector. Show that for all x, $\lambda \mapsto \mathrm{Var}(\hat{h}_\lambda(x) \,|\, X_{1:n})$ is strictly decreasing. [*Hint: Consider the eigendecomposition of* $\Phi^\top\Phi$.]
Solution: Similarly to lectures, we can obtain

$$\mathrm{Var}(\hat{h}_\lambda(x) \,|\, X_{1:n}) = \varphi(x)^\top(\Phi^\top\Phi + \lambda I)^{-1}(\Phi^\top A\Phi)(\Phi^\top\Phi + \lambda I)^{-1}\varphi(x),$$

where $A = \mathbb{E}[\{Y_{1:n} - \mathbb{E}(Y_{1:n} \,|\, X_{1:n})\}\{Y_{1:n} - \mathbb{E}(Y_{1:n} \,|\, X_{1:n})\}^\top \,|\, X_{1:n}]$. By the assumption on the variance, we can show (see lectures) that $A = \sigma^2 I$. Now considering the eigendecomposition $\Phi^\top\Phi = UDU^\top$, we have

$$(\Phi^\top\Phi + \lambda I)^{-1}(\Phi^\top\Phi)(\Phi^\top\Phi + \lambda I)^{-1} = U(D + \lambda I)^{-2}U^\top.$$

Thus

$$\mathrm{Var}(\hat{h}_\lambda(x) \,|\, X_{1:n}) = \sigma^2 \sum_{i=1}^n \frac{\{(U^\top\varphi(x))_i\}^2}{(D_{ii} + \lambda)^2},$$

which is strictly decreasing in λ.

6. In this question we investigate an alternative splitting criterion for a regression tree, based on maximising a likelihood assuming that the Y_i have a Poisson distribution. Specifically, consider the first split and where $p = 1$ with $X_1 < \cdots < X_n$. Show that

$$\max_{\gamma_L,\gamma_R} \prod_{i\leq m}(\gamma_L^{Y_i}e^{-\gamma_L}) \times \prod_{i>m}(\gamma_R^{Y_i}e^{-\gamma_R})$$

may be maximised over m with $O(n)$ computational cost.
Solution: Taking logs of the objective, we arrive at

$$\max_{\gamma_L} \sum_{i\leq m}\{Y_i \log\gamma_L - \gamma_L\} + \max_{\gamma_R} \sum_{i>m}\{Y_i \log\gamma_R - \gamma_R\}.$$

Differentiating w.r.t γ_L and γ_R, we see that the maximising quantities are A_m/m and $B_m/(n-m)$ respectively, where

$$A_m := \sum_{i\leq m} Y_i \qquad \text{and} \qquad B_m := \sum_{i>m} Y_i.$$

Thus the objective is given by

$$A_m \log(A_m/m) - A_m/m + B_m \log(B_m/(n-m)) - B_m/(n-m).$$

As $A_{m+1} = A_m + Y_{m+1}$ and $B_{m+1} = B_m - Y_{m+1}$, we see we may compute the objective at each m in $O(n)$ total time.

7. The piecewise constant function produced by a regression tree may not always approximate the underlying true regression function well. Here we imagine we have an additional univariate predictor $T_1, \ldots, T_n \in \mathbb{R}$ which we permit to contribute to the fit in a linear fashion. Specifically, consider ERM with squared error loss over class

$$\mathcal{H} := \left\{ (t, x) \mapsto t\beta + \sum_{j=1}^{J} \gamma_j \mathbb{1}_{R_j}(x) \ : \beta \in \mathbb{R}, \ \gamma \in \mathbb{R}^J \right\};$$

here the R_j are fixed (for simplicity, unlike in the case of regression trees) and partition $\mathbb{R}^p$ and moreover all $I_j := \{i : X_i \in R_j\}$ are non-empty and have been pre-computed. Assume that $T_{1:n} \in \mathbb{R}^n$ is not in the span of $\{(\mathbb{1}_{R_j}(X_i))_{i=1}^n : j = 1, \ldots, J\}$. Show that the ERM may be computed in $O(n)$ time. [*Hint: Use the matrix identity that for* $M \in \mathbb{R}^{p\times p}$, $b \in \mathbb{R}^p$ *and* $a \in \mathbb{R}$,

$$\begin{pmatrix} a & b^\top \\ b & M \end{pmatrix}^{-1} = \begin{pmatrix} s^{-1} & -s^{-1}b^\top M^{-1} \\ -s^{-1}M^{-1}b & M^{-1} + s^{-1}M^{-1}bb^\top M^{-1} \end{pmatrix},$$

where $s := a - b^\top M^{-1} b > 0$ *provided the matrix on the left is indeed invertible.*]
Solution: Note first that $J \le n$. Let $\Psi \in \mathbb{R}^{n\times J}$ have entries $\Psi_{ij} = \mathbb{1}_{R_j}(X_i)$ and let $\Phi := (T_{1:n} \ \Psi) \in \mathbb{R}^{n\times(J+1)}$. The ERM is of the form given by $\mathcal{H}$ with $(\beta, \gamma) = (\hat{\beta}, \hat{\gamma})$ satisfying

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = (\Phi^\top \Phi)^{-1} \Phi^\top Y_{1:n}.$$

Now $\Psi^\top \Psi =: D \in \mathbb{R}^{J\times J}$ is a diagonal matrix with $D_{jj} = |I_j|$ as the R_j form a partition. Also writing $b := \Psi^\top T_{1:n}$ we have $b_j = \sum_{i\in I_j} T_i$. Then each of D and b can be computed in $O(n)$ time and

$$\Phi^\top \Phi = \begin{pmatrix} a & b^\top \\ b & D \end{pmatrix}$$

where $a = \|T_{1:n}\|_2^2$. Then $s := a - b^\top D^{-1} b$ may be computed in $O(J)$ time as D is diagonal. Note also that similarly to the above, $\Psi^\top Y_{1:n}$ can be computed in $O(n)$ time, and hence also $\Phi^\top Y_{1:n} =: (u \, v) \in \mathbb{R} \times \mathbb{R}^J$. Thus using the blockwise inversion formula,

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} s^{-1}(u - b^\top D^{-1} v) \\ -us^{-1}D^{-1}b + D^{-1}v + s^{-1}(b^\top D^{-1} v) D^{-1} b, \end{pmatrix}$$

which may be computed in $O(n)$ time.

8. Consider the regression setting with squared error loss and let $\mathcal{H} = \{x \mapsto \beta^\top x : \beta \in \mathbb{R}^p\}$. Let $\Sigma_{XX} := \mathrm{Var}(X) \in \mathbb{R}^{p\times p}$ and $\Sigma_{XY} = \mathrm{Cov}(X, Y) \in \mathbb{R}^p$. Suppose Σ_{XX} is positive definite, $\mathbb{E}X = 0$ and $\mathbb{E}Y^2 < \infty$. Show that $h^* := \mathrm{argmin}_{h\in\mathcal{H}} R(h)$ is given by $h^*(x) = x^\top \beta^*$ where $\beta^* = \Sigma_{XX}^{-1}\Sigma_{XY}$.
Solution: Let $h(x) = x^\top \beta$. Then

$$R(h) = \mathbb{E}\{Y - X^\top \beta\}^2 = \mathbb{E}Y^2 - 2\Sigma_{XY}^\top \beta + \beta^\top \Sigma_{XX} \beta.$$

The above is minimised over $\beta \in \mathbb{R}^p$ by $\beta^* := \Sigma_{XX}^{-1}\Sigma_{XY}$, so $h^*(x) = x^\top \beta^*$.

9. Suppose $|\mathcal{H}|$ is finite and there exists $h^* \in \mathcal{H}$ with $R(h^*) = 0$. Show that with probability at least $1 - \delta$, every empirical risk minimiser $\hat{h}$ satisfies

$$R(\hat{h}) \le \frac{\log|\mathcal{H}| + \log(1/\delta)}{n}.$$

[*Hint:* $1 - \epsilon \le e^{-\epsilon}$.]
Solution: Let the RHS above be ϵ. Let $h \in \mathcal{H}$ be such that $R(h) > \epsilon$ (if no such h exists we are done). Then $\mathbb{P}(\ell(h(X), Y) = 1) = R(h) > \epsilon$. Note that then $\hat{R}(h) = 0$ if and only if $\ell(h(X_i), Y_i) = 0$ for all i, so $\mathbb{P}(\hat{R}(h) = 0) \le (1 - \epsilon)^n \le e^{-\epsilon n}$. Now for any ERM $\hat{h}$, $\hat{R}(\hat{h}) \le \hat{R}(h^*)$ and $R(h^*) = 0$ implies $\ell(h^*(X_i), Y_i) = 0$ almost surely, so $\hat{R}(h^*) = 0$ (almost surely). Thus

$$\begin{aligned} \mathbb{P}(R(\hat{h}) > \epsilon \text{ for some ERM } \hat{h}) &\le \mathbb{P}\left(\cup_{h:R(h)>\epsilon} \hat{R}(h) = 0\right) \\ &\le \sum_{h:R(h)>\epsilon} e^{-\epsilon n} \le |\mathcal{H}| e^{-\epsilon n} = \delta. \end{aligned}$$

10. This question is about (potentially high-dimensional) covariance matrix estimation. Suppose $Z_i \overset{\text{i.i.d.}}{\sim} N_p(0, \Sigma)$ for $i = 1, \ldots, n$ where $\Sigma \in \mathbb{R}^{p \times p}$ is a covariance matrix with $\Sigma_{jj} = 1$ for $j = 1, \ldots, p$. The maximum likelihood estimate of Σ is $\hat{\Sigma} := \frac{1}{n} \sum_{i=1}^n Z_i Z_i^\top$.

(a) Suppose V and W are mean-zero and jointly Gaussian with $\text{Var}(V) = \text{Var}(W) = 1$ and $\text{Cov}(V, W) = \rho$. Show that

$$\mathbb{E}e^{\alpha VW} = [\{1 - \alpha(1 + \rho)\}\{1 + \alpha(1 - \rho)\}]^{-1/2}$$

for $\alpha \in (-1/2, 1/2)$. [*Hint: Express VW as a difference of two independent scaled χ_1^2 random variables and use the fact that the mgf of a χ_1^2 random variable is $1/\sqrt{1 - 2\alpha}$ for $\alpha < 1/2$.*]
Solution: Note that $VW = (V + W)^2/4 - (V - W)^2/4$. $V + W$ and $V - W$ are jointly normal with 0 covariance, hence they are independent. Now $V + W \sim N(0, 2(1 + \rho))$, so when $\rho \in (-1, 1)$ $(V + W)^2/\{2(1 + \rho)\} \sim \chi_1^2$. Similarly, $(V - W)^2/\{2(1 - \rho)\} \sim \chi_1^2$. Thus (using independence)

$$\begin{aligned}\mathbb{E}(\exp(\alpha VW)) &= \mathbb{E}e^{\alpha(V+W)^2/4}\mathbb{E}e^{-\alpha(V-W)^2/4} \\ &= [\{1 - \alpha(1 + \rho)\}\{1 + \alpha(1 - \rho)\}]^{-1/2}\end{aligned}$$

provided $|(1 + \rho)\alpha/2| < 1/2$. The result is also true when $\rho = \pm 1$ as one of the terms above is simply 1. Thus the result is true for all $|\alpha| < 1/2$.

(b) Using the fact that

$$e^{-\alpha\rho}[\{1 - \alpha(1 + \rho)\}\{1 + \alpha(1 - \rho)\}]^{-1/2} \le e^{2\alpha^2}$$

whenever $|\alpha| < 1/4$ and $\rho \in [-1, 1]$, show that for fixed $j, k \in \{1, \ldots, p\}$ and $t \in (0, 1)$,

$$\mathbb{P}(|\hat{\Sigma}_{jk} - \Sigma_{jk}| \ge t) \le 2e^{-nt^2/8}.$$

Conclude that with probability at least $1 - 2/p$,

$$\max_{j,k} |\hat{\Sigma}_{jk} - \Sigma_{jk}| \le 5\sqrt{\frac{\log(p)}{n}}.$$

Solution: Suppose $j \neq k$ and let $V = Z_{1j}$ and $W = Z_{1k}$. Then $n\hat{\Sigma}_{jk}$ is a sum of i.i.d. copies of VW. Thus we have

$$\begin{aligned}\mathbb{E}\exp\{\alpha n(\hat{\Sigma}_{jk} - \Sigma_{jk})\} &= \left(\mathbb{E}e^{\alpha VW - \mathbb{E}(\alpha VW)}\right)^n \\ &\le \exp(2n\alpha^2)\end{aligned}$$

for $|\alpha| < 1/4$ using the hint. Note that this still holds when $j = k$. Now suppose $t \in (0, 1)$. Using the Chernoff bound, we get

$$\begin{aligned}\mathbb{P}(n(\hat{\Sigma}_{jk} - \Sigma_{jk}) \ge nt) &\le \inf_{0<\alpha<1/4} \exp\{n(2\alpha^2 - \alpha t)\} \\ &= e^{-nt^2/8}\end{aligned}$$

setting $\alpha = t/4$ in the last line, which is permitted since $t < 1$. The argument to bound $\mathbb{P}(n(\Sigma_{jk} - \hat{\Sigma}_{jk}) \ge nt)$ is similar, and the result follows from a union bound. Finally, we have from a union bound that

$$\mathbb{P}(\cup_{j,k} |\hat{\Sigma}_{jk} - \Sigma_{jk}| \ge t) \le 2\exp(-nt^2/8 + 2\log(p)),$$

so if $t = 5\sqrt{\log(p)/n}$, the RHS is at most $2p^{-1}$.