Marius Yamakou
Martin Betz
Sebastian Denzler

December 12, 2024
STMOL – Exercise sheet 3

This exercise sheet contains problems from Parts II and III of the course.
Please submit the exercises highlighted in red before the deadline.

Problem 1

(a) Let X be a random variable with $\mathrm{E}[X] = 0$. Suppose that the moment-generating function of X^2 is bounded at some point, that is,

$$\mathrm{E}\left[e^{X^2}\right] \leq 2.$$

Prove that X satisfies the two-sided tail bound

$$\mathrm{P}(|X| > t) \leq 2e^{(-t^2)} \text{ for all } t \geq 0.$$

(b) Prove that if X is a non-negative random variable with expectation $\mathrm{E}[X]$, then for all $t > 0$, we have $\mathrm{P}[X \geq t] \leq \mathrm{E}[X]/t$.

(c) Recall Chernoff's inequality: Let X_i be independent Bernoulli random variables with success probability p_i. Consider their sum $S_N = \sum_{i=1}^N X_i$ and denote its mean by $\mu = \mathrm{E}[S_N]$. Then, for any $t > \mu$, we have

$$\mathrm{P}(S_N \geq t) \leq e^{t-\mu}\left(\frac{\mu}{t}\right)^t.$$

Consider 200 independent coin flips. We wish to find an upper bound on the probability that the number of heads is greater or equal than 150. Use Chernoff's inequality.

(d) Let X_i, for $I = 1, \ldots, n$, be a random sample of a random variable X. Let X have mean μ and variance σ^2. Find the size of the sample (n), such that the probability that the difference between sample mean and true mean is smaller that $\frac{\sigma}{10}$ is at least 0.95. Hint: Derive a version of the Chebyshev inequality for $\mathrm{P}(|X - \mu| \geq a)$ using Markov inequality.

Solution.

(a) It holds that

$$\mathrm{P}(|X| > t) = \mathrm{P}(e^{X^2} > e^{t^2}) \leq \frac{2}{e^{t^2}}.$$

(we just apply Markov in the last step)

(b) It holds that

$$\mathrm{P}(X \geq t) = \mathrm{P}(t_{X \geq t} \geq t) = \mathrm{E}\left[_{X \geq t}\right] \leq \mathrm{E}\left[\frac{X}{t}\right] = \frac{\mathrm{E}[X]}{t}.$$

Alternative way:

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x) dx = \int_0^{\infty} x f(x) dx \geq \int_t^{\infty} x f(x) dx \geq \int_t^{\infty} t f(x) dx = t\,\mathrm{P}(X \geq t)$$

$$\implies \mathrm{E}[X] \geq t\,\mathrm{P}(X \geq t) \implies \frac{\mathrm{E}[X]}{t} \geq \mathrm{P}(X \geq t)$$

(c) Chernoff gives $e^{50}\left(\frac{2}{3}\right)^{150} = \left(\frac{8e}{27}\right)^{50}$. It is not necessary to simplify this further.

(d) Let $\hat{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then, $\mathrm{E}\left[\hat{X}\right] = \mu$ and $\mathrm{Var}\left[\hat{X}\right] = \frac{\sigma^2}{n}$. Now, we need to determine n such that

$$\mathrm{P}(|\hat{X} - \mu| \leq \frac{\sigma}{10}) \geq 0.95 \implies \mathrm{P}(|\hat{X} - \mu| \geq \frac{\sigma}{10}) \leq 0.05$$

We can write the probability as:

$$P(\sqrt{(\hat{X}-\mu)^2} \geq \frac{\sigma}{10}) = P((\hat{X}-\mu)^2 \geq \frac{\sigma^2}{100}) \leq \frac{\mathrm{Var}\left[\hat{X}\right]}{\frac{\sigma^2}{100}} = \frac{\sigma^2}{n}\frac{100}{\sigma^2} = \frac{100}{n} \leq 0.05$$

$$\implies \frac{100}{0.05} \leq n$$

Therefore, we need a sample size of $n \geq 2000$.

Problem 2

1. Estimation of diagonal covariances: Let $(X_i)_{i=1,\dots,n}$ be an i.i.d. sequence of d-dimensional vectors, drawn from a zero-mean distribution with diagonal covariance matrix $\Sigma = D$. Consider the estimate $\widehat{D} = \mathrm{diag}(\widehat{\Sigma})$, where $\widehat{\Sigma}$ is the usual sample covariance matrix. Suppose further that each component X_{ij} is sub-Gaussian with parameter at most $\sigma = 1$. Show the following:
   (a) X_{ij}^2 is sub-exponential with parameters (2,4).
   (b) $\sum_{i=1}^{n} X_{ij}^2$ is sub-exponential with parameters $(2\sqrt{n}, 4)$
   (c) For each $i = 1, \dots, d$, we get
   $$P\left(|\widehat{D}_{ii} - D_{ii}| \geq t\right) \leq 2e^{-\frac{n}{8}\min\left\{t,t^2\right\}}.$$
2. Suppose that the random vector $X \in \mathrm{R}^n$ has a $N_n(\mu, \Sigma)$ distribution, where Σ is positive. Show the the random variable $Y = (X-\mu)^T \Sigma^{-1} (X-\mu)$ is sub-exponential.

Solution.

1. (a) X_{ij}^2 is sub-exponential with parameters (2,4).

Approach 1 For this, we consider that a sub-gaussian variable of parameter at most σ will be bounded for above by a gaussian variable. Let $X \sim N(0, \sigma^2)$, and further assume $\sigma = 1$. Now, consider that X^2 follows a chi-squared distribution, and its moment generating function is defined as.

$$E\left[e^{\lambda(X^2-1)}\right] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\lambda x^2} e^{-\frac{x^2}{2}}dx = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}, \text{ for } \lambda < \frac{1}{2}$$

The moment generating function is also obtain by using the gaussian distribution and considering $E\left[X^2\right] = 1$. Following the definition of sub-exponential we have:

$$E\left[e^{\lambda(X^2-1)}\right] \leq e^{\frac{\nu^2\lambda^2}{2}} \text{ for all } \lambda^2 < \frac{1}{\alpha^2}$$

Now, considering $\nu = 2$, and $\alpha = 4$, we have that $\lambda^2 < \frac{1}{16} \implies \lambda \in (-\frac{1}{4}, \frac{1}{4})$. Therefore, the moment generation function previously calculated is bounded for these values of λ. With this it should hold that

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{\frac{2^2\lambda^2}{2}} = e^{2\lambda^2} \text{ for all } \lambda^2 < \frac{1}{16}$$

Let's focus for $\lambda \in (-\frac{1}{4}, \frac{1}{4})$. Given that all terms are positive, we can square and reorder the inequality:

$$e^{-4\lambda^2-2\lambda} \leq 1-2\lambda$$

$$\implies -4\lambda^2 - 2\lambda \leq \log 1 - 2\lambda \implies 0 \leq ln(1-2\lambda) + 4\lambda^2 + 2\lambda = f(\lambda)$$

It easy to show that $f(x)$ is a convex function in the domain of lambda. Therefore we can calculate its minimum with first and second order condition. If $\min f(\lambda) \geq 0$ for $|\lambda| < \frac{1}{4}$, the inequality holds and the variable is sub-exponential(2,4).

$$FOC: \frac{df(\lambda)}{d\lambda} = -\frac{2}{1-2\lambda} + 8\lambda + 2 = -2 + 8\lambda - 16\lambda^2 + 2 - 4\lambda = 0$$

$$4\lambda(1-4\lambda) = 0 \implies \lambda = 0 \vee \lambda = 1/4, \text{ we can see that } 1/4 \text{ is not a minimizer.}$$

The second derivative evaluated in $\lambda = 0$ has a value of 4, therefore $\lambda = 0$ is a proper minimizer of $f(\lambda)$, and $f(0) = 0$. Thus the inequality holds and the variable X^2 is sub-exponential of parameter (2,4).

Possible Alternative Another way to approach the problem will be: Let $Z = X^2 - \mathrm{E}[X^2]$. Then, we calculated its moment generation function using the Taylor expansion of the exponential,

$$\mathrm{E}\left[e^{\lambda Z}\right] \leq \mathrm{E}\left[1 + \sum_{k=1}^{\infty} \frac{\lambda^k Z^k}{k!}\right]$$

Following this, we can keep bounding using Jensen's inequality and the bounds available given that X is sub-gaussian the parameter at most 1.

(b) Let $Z_{ij} = X_{ij}^2$, and therefore be sub-exponential with parameters $(2, 4)$. Now we compute the moment generating function:

$$\mathrm{E}\left[e^{\lambda \sum_{i=1}^{n}(Z_{ij} - \mathrm{E}[Z_{ij}])}\right] = \prod_{i=1}^{n} \mathrm{E}\left[e^{\lambda(Z_{ij} - \mathrm{E}[Z_{ij}])}\right]$$

Now, following the bounds obtained beforehand:

$$\prod_{i=1}^{n} \mathrm{E}\left[e^{\lambda(Z_{ij} - \mathrm{E}[Z_{ij}])}\right] \leq \prod_{i=1}^{n} e^{\nu^2 \frac{\lambda^2}{2}} = e^{\sum_{i=1}^{n}(\nu^2 \frac{\lambda^2}{2})} \qquad \forall |\lambda| \leq \frac{1}{4}$$

$$\implies E\left[e^{\lambda \sum_{i=1}^{n}(Z_{ij} - \mathrm{E}[Z_{ij}])}\right] \leq e^{(\sqrt{n}\nu)^2 \frac{\lambda^2}{2}} \qquad \forall |\lambda| \leq \frac{1}{4}$$

Finally, we can conclude that $\sum_{i=1}^{n} X_{ij}^2$ is sub-exponential with parameters $(2\sqrt{n}, 4)$.

(c) $\widehat{D}_{ii}$ is the usual sample covariance matrix, and it's defined as $\widehat{D}_{ii} = \dfrac{1}{n}\sum_{i=1}^{n} x_{ij}^2$. This estimator is unbiased, i.e., $\mathrm{E}\left[\widehat{D}_{ii}\right] = D$. Also, following the previous exercises we have that $\widehat{D}_{ii}$ is sub-exponential de parameters $(\frac{2}{\sqrt{n}}, \frac{4}{n})$. Now, given sub-exponential concentration

$$\mathrm{P}\left(|\widehat{D}_{ii} - D_{ii}| \geq t\right) \leq 2\exp^{-\frac{1}{2}\min\left\{\frac{t}{\alpha}, \frac{t^2}{\nu^2}\right\}}.$$

Replacing *nu* and α for their respective values, we get:

$$\mathrm{P}\left(|\widehat{D}_{ii} - D_{ii}| \geq t\right) \leq 2e^{-\frac{n}{8}\min\{t, t^2\}}.$$

2. $Y = (X-\mu)^T \Sigma^{-1} (X-\mu)$. Let's consider the spectral decomposition of $\Sigma = Q\Lambda Q^T$, where $Q^T Q = I$. Now, $\Sigma^{-\frac{1}{2}}$ is ten defined as $Q\Lambda^{\frac{1}{2}}Q^T$, and therefore $\Sigma^{-1} = \Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}$. Let $Z = \Sigma^{-\frac{1}{2}}(X-\mu)$, and corresponds to random variable that follows a normal standard distribution. Then,

$$Y = Z^T Z = \sum_i Z_i^2$$

Therefore, as presented in the lecture, $Y \sim \chi^2(n)$ is a sub-exponential variable.

Problem 3

Convexity and norms: A norm $\|\cdot\|$ over $\mathbb{R}^n$ is defined by the properties:

(i) non-negativity: $\|x\| \geq 0$ for all $x \in \mathbb{R}^n$ with equality if and only if $x = 0$,

(ii) absolute scalability: $\|a\,x\| = |a|\,\|x\|$ for all $a \in \mathbb{R}$ and $x \in \mathbb{R}^n$,

(iii) triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^n$.

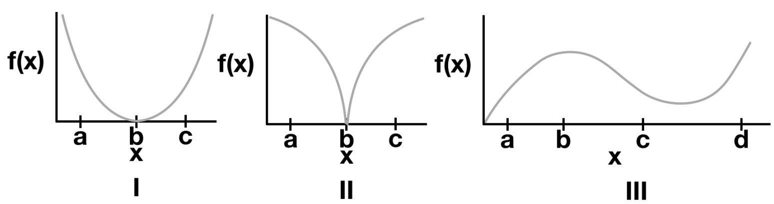
(a) Show that $\|x\|_1 = \sum_{i=1}^n |x_i|$ is a norm. (Hint: for (*iii*), begin by showing that $|a + b| \leq |a| + |b|$ for all $a, b \in \mathbb{R}$.) (Correspond to the penalty for LASSO.)

(b) Show that $f(x) = \left(\sum_{i=1}^n |x_i|^{1/2}\right)^2$ is not a norm. (Hint: it suffices to find two points in $n = 2$ dimensions such that the triangle inequality does not hold.)

(c) Show that $\|x\|_2 := \left(\sum_{i=1}^n |x_i|^2\right)^{1/2}$ is a norm. (Correspond to the penalty for ridge regression.)

We say a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex on a set A if $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$ for all $x, y \in A$ and $\lambda \in [0, 1]$.

(d) For each of the functions below (I-III), state whether each one is convex on the given interval or state why not with a counterexample using any of the points a, b, c, d in your answer.



1 Function in panel I on $[a, c]$
2 Function in panel II on $[a, c]$
3 Function in panel III on $[a, d]$
4 Function in panel III on $[c, d]$

(e) For $i = 1, \ldots, n$ let $\ell_i(w)$ be convex functions over $w \in \mathbb{R}^d$ (e.g., $\ell_i(w) = (y_i - w^\top x_i)^2$), $\|\cdot\|$ is any norm, and $\lambda > 0$. Show that

$$\sum_{i=1}^n \ell_i(w) + \lambda \|w\|$$

is convex over $w \in \mathbb{R}^d$ (Hint: Show that if f, g are convex functions, then $f(x) + g(x)$ is also convex.)

Solution.

(a) We show that (i)-(iii) hold.

(i) $\sum_{i=1}^n |x_i| \geq 0$; $x = 0 \Rightarrow \sum_{i=1}^n |x_i| = 0$, $\sum_{i=1}^n |x_i| = 0 \Rightarrow x_i = 0$ for all $i \in \{1, \ldots, n\}$.

(ii) $\|ax\|_1 = \sum_{i=1}^n |ax_i| = |a| \sum_{i=1}^n |x_i| = |a| \|a\|_1$.

(iii) $\|x + y\|_1 = \sum_{i=1}^n |x_i + y_i| \leq \sum_{i=1}^n |x_i| + \sum_{i=1}^n |y_i| = \|x\| + \|y\|$.

(b) We choose the points $x = (0,4)^T$ and $y = (4,0)^T$. For these points, we obtain

$$f(x+y) = \Big(\sum_{i=1}^{n} |x_i + y_i|^{12}\Big)^2 = \big(4^{12} + 4^{12}\big)^2 = 16$$
$$f(x) + f(y) = (4^{12})^2 + (4^{12})^2 = 8.$$

Hence, the triangle inequality does not hold and $f(\cdot)$ cannot be a norm.

(c) We show that (i)-(iii) hold.

(i) $x_2 \geq 0$; $x = 0 \Rightarrow 0_2 = \big(\sum_{i=1}^n |0|^2\big)^{12} = 0$, $0_2 = 0 \Rightarrow \big(\sum_{i=1}^n |x_i|^2\big)^{12} = 0 \Rightarrow \sum_{i=1}^n |x_i|^2 = 0 \Rightarrow x_i = 0$ for all $i \in \{1, \dots, n\}$.

(ii) $\|ax\|_2 = \big(\sum_{i=1}^n |ax_i|^2\big)^{12} = |a|\big(\sum_{i=1}^n |x_i|^2\big)^{12} = |a|\|a\|_2$.

(iii) $\|x+y\|_2 = \big(\sum_{i=1}^n |x_i + y_i|^2\big)^{12} = \big(\sum_{i=1}^n {x_i}^2 + 2\sum_{i=1}^n x_i y_i + \sum_{i=1}^n {y_i}^2\big)^{12} = \big(\sum_{i=1}^n {x_i}^2 + 2x^T y + \sum_{i=1}^n {y_i}^2\big)^{12} = \big(x_2^2 + 2x^T y + y_2^2\big)^{12} \leq \big(x_2^2 + 2x_2 y_2 + y_2^2\big)^{12} = \big((x_2 + y_2)^2\big)^{12} = x_2 + y_2$, where we used the Cauchy-Schwarz inequality.

(d)
(a) convex
(b) not convex on $[b, c]$: Every function value in (b, c) is greater than the corresponding convex combination of the function values $f(b)$ and $f(c)$.
(c) not convex on $[a, c]$: Every function value in (a, c) is greater than the corresponding convex combination of the function values $f(a)$ and $f(c)$.
(d) convex

(e) Let f and g be two convex functions. Then, for every $\lambda \in [0,1]$,

$$\begin{aligned}(f+g)(\lambda x + (1-\lambda)y) &= f(\lambda x + (1-\lambda)y) + g(\lambda x + (1-\lambda)y) \\ &\leq \lambda f(x) + (1-\lambda) f(y) + \lambda g(x) + (1-\lambda) g(y) \\ &= \lambda(f(x) + g(x))(1-\lambda)(f(y) + g(y)) \\ &= \lambda(f+g)(x)(1-\lambda)(f+g)(y),\end{aligned}$$

which shows that $f + g$ is convex. With this, it is now sufficient to show that every norm is a convex function:

$$\lambda x + (1-\lambda)y \overset{(iii)}{\leq} \lambda x + (1-\lambda)y \overset{(ii)}{=} \lambda x + (1-\lambda)y.$$

Problem 4

The following questions should be answered without referring to external materials. Briefly justify your answers with a few words.

(a) How does lasso regression differ from ridge regression?

(b) Why do least squares fail in high dimensions?

(c) In a LASSO regression, if the regularization parameter λ is very high, what happens to the estimated regression coefficients?

(d) True or False: The LASSO is a convex optimization problem.

Solution.

(a) LASSO regression can achieve feature selection (by setting feature weights to exactly zero) while ridge regression cannot.

(b) $X'X$ has no longer full rank when $p > n$. The OLS results in infinitely many solutions, leading to over-fitting in HD.

(c) The model can shrink the coefficients of uninformative features to exactly zero.

(d) True (see Problem 3).

Problem 5

Ridge Regression: Consider the linear regression model

$$y = X\beta_0 + \varepsilon$$

with $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times d}$ and $\varepsilon \in \mathbb{R}^n$ some random noise vector. The ridge regression estimator is employed when $\mathrm{rk}(X'X) < d$. It is defined for a given parameter $\lambda > 0$ by

$$\widehat{\beta} = \mathrm{argmin}_{\beta \in \mathbb{R}^d} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2. \tag{1}$$

(a) Show that for any $\lambda > 0$ the solution to the minimization problem (1) is

$$\widehat{\beta} = (X'X + \lambda I_{d \times d})^{-1} X'y.$$

You may use that $\frac{\partial}{\partial \beta} \beta' X' X \beta = 2X'X\beta$. Note also that you need to argue why $X'X + \lambda I_{d \times d}$ is invertible.

(b) Compute the bias $\mathrm{E}(\widehat{\beta}) - \beta_0$.

Solution.

(a) We take the derivative with respect to β and set it to zero.

$$\begin{aligned}
&\frac{\partial}{\partial \beta}(y - X\beta_2^2 + \lambda\beta_2^2) \\
=&\frac{\partial}{\partial \beta}((y - X\beta)'(y - X\beta) + \lambda\beta'\beta) \\
=&\frac{\partial}{\partial \beta}(y'y - \beta'X'y - y'X\beta + \beta'X'X\beta + \lambda\beta'\beta) \\
=&\frac{\partial}{\partial \beta}(-2\beta'X'y + \beta'X'X\beta + \lambda\beta'\beta) \\
=&-2X'y + 2X'X\beta + 2\lambda I\beta \overset{!}{=} 0 \\
\Rightarrow&(X'X + \lambda I)\beta = X'y \\
\Rightarrow&\hat{\beta} = (X'X + \lambda I)^{-1}X'y.
\end{aligned}$$

(b)

$$
\begin{aligned}
E(\hat{\beta}) - \beta_0 =& (X'X + \lambda I)^{-1} X' E(y) - \beta_0 \\
=& (X'X + \lambda I)^{-1} X' E(X\beta_0 + \varepsilon) - \beta_0 \\
\overset{E(\varepsilon)=0}{=}& (X'X + \lambda I)^{-1} X' X \beta_0 - \beta_0 \\
=& (X'X + \lambda I)^{-1} (X'X - (X'X + \lambda I))\beta_0 \\
=& (X'X + \lambda I)^{-1} (-\lambda I)\beta_0 \\
=& -\lambda (X'X + \lambda I)^{-1} \beta_0.
\end{aligned}
$$

Problem 6

Consider the sub-Gaussian sequence model

$$Y = \theta + \sigma\varepsilon,$$

where $\varepsilon \in \mathbb{R}^n$ consists of independent mean-zero 1-sub-Gaussian components, $\theta \in \mathbb{R}^n$ and $\sigma > 0$. The soft-thresholding operator, defined for $v \in \mathbb{R}$ by

$$
S_\lambda(v) = \begin{cases} v - \lambda, & v > \lambda, \\ 0, & |v| \leq \lambda, \\ v + \lambda, & v < -\lambda \end{cases}
$$

gives the soft-thresholding estimator (when applied elementwise) $\widehat{\theta} := S_\lambda(Y)$. We suppose further that θ is s-sparse, meaning that $\|\theta\|_0 = \sum_{j=1}^n 1_{\{\theta_j \neq 0\}} = s$.

(a) Show that if $\lambda \geq \sigma\|\varepsilon\|_{\max}$, then

$$\|\widehat{\theta} - \theta\|_2^2 \leq 4s\lambda^2$$

(b) Show that

$$\mathrm{P}\left(\|\widehat{\theta} - \theta\|_2 > 2\sqrt{s}\lambda\right) \leq \frac{1}{2n}$$

for $\lambda = 2\sqrt{\sigma^2 \log(2n)}$.

Solution.

(a)

$$
\begin{aligned}
\hat{\theta}_j - \theta_j = S_\lambda(Y_j) - \theta_j &= \begin{cases} Y_j - \lambda - \theta_i & Y_j > \lambda \\ -\theta_j & Y_j \in [-\lambda, \lambda] \\ Y_j + \lambda - \theta_j & Y_j < -\lambda \end{cases} \\
&= \begin{cases} \sigma\varepsilon_j - \lambda & Y_j > \lambda \\ -\theta_i & Y_j \in [-\lambda, \lambda] \\ \sigma\varepsilon_j + \lambda & Y_j < -\lambda \end{cases}
\end{aligned}
$$

$$\begin{aligned}\hat{\theta} - \theta_2^2 &= \sum_{j=1}^{n} \Big((\sigma\varepsilon_j - \lambda)\mathbb{1}_{\{\sigma\varepsilon_j + \theta_j > \lambda\}} - \theta_j \mathbb{1}_{\{\sigma\varepsilon_j+\theta_j \in [-\lambda,\lambda]\}} + (\sigma\varepsilon_j + \lambda)\mathbb{1}_{\{\sigma\varepsilon_j + \theta_j < -\lambda\}} \Big)^2 \\ &= \sum_{j=1}^{n} \Big((\sigma\varepsilon_j - \lambda)^2\mathbb{1}_{\{\sigma\varepsilon_j + \theta_j > \lambda\}} + \theta_j^2 \mathbb{1}_{\{\sigma\varepsilon_j+\theta_j \in [-\lambda,\lambda]\}} + (\sigma\varepsilon_j + \lambda)^2\mathbb{1}_{\{\sigma\varepsilon_j + \theta_j < -\lambda\}} \Big),\end{aligned}$$

where for $\theta_j = 0$, the summand is equal to 0, since $\sigma\varepsilon_j \leq \max|\sigma\varepsilon_j| \leq \lambda$ by assumption, so that, assuming that the $\theta_j \neq 0$ for the first s indices, we obtain

$$= \sum_{j=1}^{s} 4\lambda^2 = s4\lambda^2,$$

since

$$\begin{aligned}(\sigma\varepsilon_j - \lambda)^2 &= \sigma^2\varepsilon_j^2 + \lambda^2 - 2\lambda\sigma\varepsilon_j \\ &\leq \sigma^2\varepsilon^2 + \lambda^2 + 2\lambda\sigma\varepsilon \\ &\leq \lambda^2 + \lambda^2 + 2\lambda^2 = 4\lambda^2.\end{aligned}$$

The computation works analogously for $(\sigma\varepsilon_j + \lambda)^2$.

(b)

$$\begin{aligned}P(\hat{\theta} - \theta_2 > 2\lambda\sqrt{s}) &= P(\hat{\theta} - \theta_2 > 4\lambda^2 s) \\ &\leq P(\{\lambda < \sigma|\varepsilon_1|\} \cup \ldots \cup \{\lambda < \sigma|\varepsilon_n|\}) \\ &\leq \sum_{i=1}^{n} P(\lambda < \sigma|\varepsilon_i|) \\ &= \sum_{i=1}^{n} P(|\varepsilon_i| > \frac{1}{\sigma}) \\ &\leq \sum_{i=1}^{n} 2e^{-(\frac{\lambda}{\sigma})^2 2} \\ &= n2e^{-(\frac{\lambda}{\sigma})^2 2} = e^{\log(2n) - -(\frac{\lambda}{\sigma})^2 2},\end{aligned}$$

where the first inequality holds due to a): $\lambda \geq \sigma\varepsilon_{\max} \Rightarrow \hat{\theta} - \theta_2^2 \leq 4s\lambda^2$ or $\{\lambda \geq \sigma\varepsilon_{\max}\} \subset \{\hat{\theta} - \theta_2^2 \leq 4s\lambda^2\}$. Note that with $\{\lambda \geq \sigma\varepsilon_{\max}\} = \{\lambda \geq \sigma|\varepsilon_1|\} \cap \ldots \cap \{\lambda \geq \sigma|\varepsilon_n|\}$. it is also true that $\{\hat{\theta} - \theta_2^2 \leq 4s\lambda^2\}^C \subset \{\lambda \geq \sigma\varepsilon_{\max}\}^C = \big(\{\lambda \geq \sigma|\varepsilon_1|\} \cap \ldots \cap \{\lambda \geq \sigma|\varepsilon_n|\}\big)^C = \{\lambda \geq \sigma|\varepsilon_1|\} \cup \ldots \cup \{\lambda \geq \sigma|\varepsilon_n|\}$.
We obtain

$$P(\hat{\theta} - \theta_2 > 2\sqrt{s}\sqrt{\log(2n)\sigma^2}) \leq e^{\log(2n) - 2\log(2n)} = e^{-\log(2n)} = \frac{1}{2n}.$$

Problem 7

For the orthogonal case, i.e. when $X'X = I$, derive the following explicit forms for estimators,

(a) For ridge:

$$\widehat{\beta}^{Ridge} = \widehat{\beta}^{OLS}/(1+\lambda).$$

(b) For lasso:

$$\widehat{\beta}_i^{Lasso} = \text{sign}(\widehat{\beta}_i^{OLS})(|\widehat{\beta}_i^{OLS}| - \lambda)_+,$$

where $\widehat{\beta}^{OLS}$ is the regular OLS estimator and $\widehat{\beta}_i^{OLS}$ its ith component. Note that the results can differ depending on how one chooses the multiplicative constants. The solutions in Problem 7 are based on the following objective functions:

$$\widehat{\beta}^{Ridge} = \left\{ \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

$$\widehat{\beta}^{LASSO} = \left\{ \frac{1}{2} \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

Solution.

(a) For ridge regression, we know

$$\begin{aligned}\widehat{\beta}^{Ridge} &= \left(X^T X + \lambda I\right)^{-1} X^T y \\ &= \frac{1}{1+\lambda} X^T y = \frac{\widehat{\beta}^{OLS}}{(1+\lambda)}.\end{aligned}$$

(b) For lasso let us write the objective with matrices:

$$\widehat{\beta}^{LASSO} = \left\{ \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1 \right\}$$

$$= \left\{ \frac{1}{2}(y^t y - 2y^T X\beta + \beta^T X^T X\beta) + \lambda||\beta||_1 \right\} \equiv \left\{ -y^T X\beta + \frac{1}{2}\beta^T\beta + \lambda||\beta||_1 \right\}$$

$$= \left\{ \lambda||\beta||_1 - \beta^T \widehat{\beta}^{OLS} + \frac{1}{2}\beta^T\beta \right\}$$

$$= \left\{ \sum_i \lambda|\beta_i| - \beta_i \widehat{\beta}_i^{OLS} + \frac{1}{2}\beta_i^2 \right\}.$$

We can see that the problem is separable, thus it can be solve for each individual i separately. We have two cases:

- When $\widehat{\beta}_i^{OLS} \geq 0$, we have that the optimal solution follows $\beta_i^\star \geq 0$. It can be show that if $\beta^\star < 0$, the exist a new solution within an ε-neighborhood of $\beta^\star$ with better objective, contradicting the optimality of $\beta^\star$. Thus, the problem to solve is reduced to

$$\min_{\beta \geq 0} \left\{ \beta_i(\lambda - \widehat{\beta}_i^{OLS}) + \frac{1}{2}\beta_i^2 \right\}.$$

  And it's optimal value is achieved when $\beta^\star = \widehat{\beta}_i^{OLS} - \lambda$. However, as $\beta \geq 0$ we need to define the solution fo only when $\beta^\star$ is non-negative, i.e., $\beta^\star = (\widehat{\beta}_i^{OLS} - \lambda)_+$.

- Analogously, when $\widehat{\beta}_i^{OLS} \geq 0$, we have that the optimal solution follows $\beta_i^\star \leq 0$. Then, we now solve

$$\min_{\beta \leq 0} \left\{ -\beta_i(\lambda + \widehat{\beta}_i^{OLS}) + \frac{1}{2}\beta_i^2 \right\}.$$

  which has solution $\beta^\star = (\widehat{\beta}_i^{OLS} + \lambda)_-$.

In both cases the solution ca be written as:

$$\widehat{\beta}_i^{Lasso} = \text{sign}(\widehat{\beta}_i^{OLS})(|\widehat{\beta}_i^{OLS}| - \lambda)_+,$$

Problem 8

Consider the linear regression problem

$$y = X\beta + \varepsilon,$$

with $y \in \mathbb{R}^n$, $X = (x_{ij})_{i=1,\dots,n;j=1,\dots,p} \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$, and $\varepsilon \in \mathbb{R}^n$.

Suppose we have an orthogonal design matrix, i.e. $X^T X = I_{p\times p}$.

(a) Write down the classical ordinary least squares estimator under the assumption of an orthogonal design matrix. Denote each component of the vector as $\widehat{\beta}_i^{OLS}$, with $i = 1, \dots, p$.

(b) Then, the Ridge regression problem can be written as

$$\sum_{i=1}^{p} (\beta_i - \widehat{\beta}_i^{OLS})^2 + \lambda \sum_{i=1}^{p} \beta_i^2.$$

Derive the Ridge regression estimator for the i-th component in terms of $\widehat{\beta}_i^{OLS}$.

Solution.

(a) Since we know that the OLS estimator is $\widehat{\beta}_{OLS} = (X^T X)^{-1} X^T y$, and since $(X^T X)^{-1}$ is the unit matrix, this simplifies to $\widehat{\beta}_{OLS} = X^T y$. Then,

$$\widehat{\beta}_i^{OLS} = \sum_{j=1}^{n} x_{ji} y_j \ .$$

(b) By taking the first derivative with respect to β_i, we get

$$\widehat{\beta}_i^{RIDGE} = \frac{1}{\lambda + 1} \widehat{\beta}_i^{OLS} \ .$$

Note why we can write ridge problem in terms of OLS when we have orthogonal design:

$$\begin{aligned}
(y - X\beta)'(y - X\beta) &= y'y + \beta'\beta - 2y'X\beta \\
&= y'y + \beta'\beta - 2\widehat{\beta}'_{OLS}\beta \\
&= y'y + \beta'\beta - 2\widehat{\beta}'_{OLS}\beta + \widehat{\beta}'_{OLS}\widehat{\beta}_{OLS} - \widehat{\beta}'_{OLS}\widehat{\beta}_{OLS} \\
&= y'y + \beta'\beta - 2\widehat{\beta}'_{OLS}\beta + \widehat{\beta}'_{OLS}\widehat{\beta}_{OLS} - y'XX'y \\
&= (\widehat{\beta}_{OLS} - \beta)'(\widehat{\beta}_{OLS} - \beta) + y'(I - XX')y
\end{aligned}$$

Since the last term is independent of β, we only need to consider the first term.

Problem 9

Recall the Ridge regression optimization problem

$$\widehat{\beta} = \min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

Considering that $n = 2$, $p = 2$, $x_{11} = x_{12}, x_{21} = x_{22}$.

(a) Write down the Ridge regression optimization problem in the described setting, and simplify it as much as possible.

(b) Show that, in this setting, the Ridge coefficient estimators satisfy $\widehat{\beta}_1 = \widehat{\beta}_2$.

(c) Write out the LASSO optimization problem in this setting.

Solution.

a)

$$\begin{aligned}
&\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \\
=&\sum_{i=1}^{n} \left(y_i - \beta_1 x_{i1} - \beta_2 x_{i2} \right)^2 + \lambda(\beta_1^2 + \beta_2^2) \\
=&(y_1 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_1 x_{21} - \beta_2 x_{22})^2 + \lambda(\beta_1^2 + \beta_2^2) \\
=&(y_1 - \beta_1 x_{11} - \beta_2 x_{11})^2 + (y_2 - \beta_1 x_{22} - \beta_2 x_{22})^2 + \lambda(\beta_1^2 + \beta_2^2).
\end{aligned}$$

$$\implies \widehat{\beta} = \min_{\beta_1, \beta_2} \left\{ (y_1 - \beta_1 x_{11} - \beta_2 x_{11})^2 + (y_2 - \beta_1 x_{22} - \beta_2 x_{22})^2 + \lambda(\beta_1^2 + \beta_2^2) \right\}.$$

b)

$$\begin{aligned}
&\frac{d}{d\beta_1} [(y_1 - \beta_1 x_{11} - \beta_2 x_{11})^2 + (y_2 - \beta_1 x_{22} - \beta_2 x_{22})^2 + \lambda(\beta_1^2 + \beta_2^2)] \\
=&2(y_1 - \beta_1 x_{11} - \beta_2 x_{11})(-x_{11}) + 2(y_2 - \beta_1 x_{22} - \beta_2 x_{22})(-x_{22}) + 2\lambda\beta_1 \\
=&2[-y_1 x_{11} - y_2 x_{22} + \beta_1(x_{11}^2 + x_{22}^2) + \beta_2(x_{11}^2 + x_{22}^2) + \lambda\beta_1].
\end{aligned}$$

$$\begin{aligned}
&\frac{d}{d\beta_2} [(y_1 - \beta_1 x_{11} - \beta_2 x_{11})^2 + (y_2 - \beta_1 x_{22} - \beta_2 x_{22})^2 + \lambda(\beta_1^2 + \beta_2^2)] \\
=&2(y_1 - \beta_1 x_{11} - \beta_2 x_{11})(-x_{11}) + 2(y_2 - \beta_1 x_{22} - \beta_2 x_{22})(-x_{22}) + 2\lambda\beta_2 \\
=&2[-y_1 x_{11} - y_2 x_{22} + \beta_1(x_{11}^2 + x_{22}^2) + \beta_2(x_{11}^2 + x_{22}^2) + \lambda\beta_2].
\end{aligned}$$

$$\begin{aligned}
&2[-y_1 x_{11} - y_2 x_{22} + \beta_1(x_{11}^2 + x_{22}^2) + \beta_2(x_{11}^2 + x_{22}^2) + \lambda\beta_2] \stackrel{!}{=} 0 \\
&2[-y_1 x_{11} - y_2 x_{22} + \beta_1(x_{11}^2 + x_{22}^2) + \beta_2(x_{11}^2 + x_{22}^2) + \lambda\beta_1] \stackrel{!}{=} 0 \\
&\implies \beta_1 = \beta_2.
\end{aligned}$$

c) $\widehat{\beta} = \min_{\beta_1, \beta_2} \left\{ (y_1 - \beta_1 x_{11} - \beta_2 x_{11})^2 + (y_2 - \beta_1 x_{22} - \beta_2 x_{22})^2 + \lambda|\beta_1| + \lambda|\beta_2| \right\}.$

Problem 10

The LASSO problem is not always strictly convex, and thus does not necessarily have a unique solution according to standard convexity theory. However, we can define a modified problem, known as the elastic-net optimization problem, that is always strictly convex:

$$\min_{\beta} \left(\|y - X\beta\|_2^2 + \alpha_1 \|\beta\|_2^2 + \alpha_2 \|\beta\|_1 \right) \tag{2}$$

where α_1, α_2 are nonnegative tuning parameters. Besides ensuring uniqueness for all X, the elastic-net combines some of the desirable predictive characteristics of Ridge regression with the sparsity properties of LASSO.

Show how one can turn this into a LASSO problem, using an augmented version of X and y.

Solution.

Consider $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, and $\beta \in \mathbb{R}^p$.

- First, let $\overline{X}$ be an augmented version of X defined as:

  $$\overline{X} = \begin{pmatrix} X \\ \gamma I_{p\times p} \end{pmatrix}.$$

  Following this, we augment y by a zero vector of dimension p, i.e.,

  $$\overline{y} = \begin{pmatrix} y \\ 0_{p\times 1} \end{pmatrix}.$$

  Then we have

  $$\|\overline{y} - \overline{X}\beta\|_2^2 = \left\| \begin{pmatrix} y - X\beta \\ 0_p - \gamma I_{p\times p}\beta \end{pmatrix} \right\|_2^2 = \left\| \begin{pmatrix} y - X\beta \\ -\gamma\beta \end{pmatrix} \right\|_2^2 = \|y - X\beta\|_2^2 + \gamma^2 \|\beta\|_2^2 \tag{3}$$

- Consider now the LASSO problem for $\overline{y}$ and $\overline{X}$

  $$\min_{\beta} \|\overline{y} - \overline{X}\beta\|_2^2 + \alpha_2 \|\beta\|_1$$

  which, by making use of (3), is

  $$\min_{\beta} \|y - X\beta\|_2^2 + \gamma^2 \|\beta\|_2^2 + \alpha_2 \|\beta\|_1$$

  Therefore, by choosing $\gamma = \sqrt{\alpha_1}$ we get the original problem.

Problem 11

Consider a linear regression problem where $p >> n$, and assume that the rank of X is n. Let the SVD of $X = UDV^T = RV^T$, where R is $n \times n$ nonsingular, and V is $p \times n$ with orthonormal columns.

(a) Show that there are infinitely many least-squares solutions all with zero residuals.

(b) Show that the Ridge-regression estimate for β can be written as

$$\widehat{\beta}_\lambda = V(R^T R + \lambda I)^{-1} R^T y. \tag{4}$$

(c) Show that when $\lambda = 0$, the solution $\widehat{\beta}_0 = VD^{-1}U^T y$ has zero residuals, and is unique in that it has the smallest Euclidean norm among all zero-residual solutions.

Solution.

(a) Since $X \in \mathbb{R}^{n\times p}$ has rank $n \le p$, then exists $v \in \mathbb{R}^p \neq 0$ such that $Xv = 0$. Let $\widehat{\beta}$ be a zero residual solution for the least-squares problems, i.e., $\widehat{\beta} = \min\limits_{\beta} \left\{ \|y - X\beta\|_2^2 \right\}$. Then, for every $k \in \mathbb{R}$, we have :

$$\|y - X\left(\widehat{\beta} + kv\right)\|_2^2 = \|y - X\widehat{\beta} - Xkv\|_2^2 \equiv \|y - X\widehat{\beta}\|_2^2.$$

Therefore, there are infinitely many least-squares solutions all with zero residuals.

(b) We know that the Ridge-regression estimator β is computed as:

$$\beta = \left(X^T X + \lambda I\right)^{-1} X^T y$$

Therefore, it solves the equation $X^T(y - X\beta) = \lambda\beta$. Then, by (4), $X^T\left(y - X\widehat{\beta}_\lambda\right) = \lambda\widehat{\beta}_\lambda$. Working on the left side of this last equation, we have:

$$\begin{aligned}
X^T\left(y - X\widehat{\beta}_\lambda\right) &= X^T\left(y - XV(R^TR + \lambda I)^{-1}R^Ty\right) \\
&\overset{X=RV^T}{=} VR^T\left(y - RV^TV(R^TR + \lambda I)^{-1}R^Ty\right) \\
&\overset{V^TV=I}{=} VR^T\left(y - R(R^TR + \lambda I)^{-1}R^Ty\right) \\
&= V\left(R^Ty - R^TR(R^TR + \lambda I)^{-1}R^Ty\right) \\
&= V\left(I - R^TR(R^TR + \lambda I)^{-1}\right)R^Ty \\
&= V\left(I - \left(R^TR + \lambda I - \lambda I\right)\left(R^TR + \lambda I\right)^{-1}\right)R^Ty \\
&= V\left(I - \left(R^TR + \lambda I\right)\left(R^TR + \lambda I\right)^{-1} + (\lambda I)\left(R^TR + \lambda I\right)^{-1}\right)R^Ty \\
&= V\left(I - I + \lambda I\left(R^TR + \lambda I\right)^{-1}\right)R^Ty \\
&= V\left(\lambda I\left(R^TR + \lambda I\right)^{-1}\right)R^Ty \\
&= \lambda V\left(\left(R^TR + \lambda I\right)^{-1}\right)R^Ty \\
&= \lambda\widehat{\beta}_\lambda
\end{aligned}$$

Therefore, the Ridge-regression estimate for β can be written as

$$\widehat{\beta}_\lambda = V(R^TR + \lambda I)^{-1}R^Ty.$$

(c)
- Zero residual implies that $y = X\beta$.

$$
\begin{aligned}
X\widehat{\beta}_0 &= XVD^{-1}U^Ty \\
&= UDV^TVD^{-1}U^Ty \\
&= UDD^{-1}U^Ty \\
&= UU^Ty \\
&= y.
\end{aligned}
$$

Then, $\widehat{\beta}_0$ has zero residual.

- If the solution is not unique, we can construct a zero residual solution as follows:

$$\beta = \widehat{\beta}_0 + v,$$

with $v \in \mathbb{R}^p \neq 0$. For β to be zero residual it needs to satisfy $X\beta = y$, i.e, $X\left(\widehat{\beta}_0 + v\right) = y$. Given that $\widehat{\beta}_0$ already has zero residual, we have

$$Xv = RV^Tv = 0.$$

Taking into consideration that R is $n \times n$ nonsingular, we have then that $V^Tv = 0$. Now, taking the Euclidean norm (squared) of β, we have

$$
\begin{aligned}
\|\beta\|_2^2 &= (\widehat{\beta}_0 + v)^T(\widehat{\beta}_0 + v) && (5)\\
&= \widehat{\beta}_0^T\widehat{\beta}_0 + v^Tv + 2\widehat{\beta}_0^Tv && (6)\\
&= \widehat{\beta}_0^T\widehat{\beta}_0 + v^Tv + 2y^TUD^{-1}V^Tv && (7)\\
&= \widehat{\beta}_0^T\widehat{\beta}_0 + v^Tv + 0 && (8)\\
&= \|\widehat{\beta}_0\|_2^2 + v^Tv. && (9)
\end{aligned}
$$

Finally, since $v^Tv > 0$ we have that $\|\beta\|_2^2 > \|\widehat{\beta}_0\|_2^2$, i.e., $\widehat{\beta}_0$ has the smallest Euclidean norm.