1. Consider the Lasso problem \( \min_{\beta} \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \), where \( \lambda > 0 \) is a tuning parameter.

(a) Let \( \hat{\beta} \) be a minimizer of the Lasso problem with \( j^{th} \) componenet \( \hat{\beta}_j \). Denote \( X_j \) to be the \( j \)-th column of \( X \). Show that
\[
\begin{cases}
  \lambda = n^{-1}X_j^T(Y - X\hat{\beta}) & \text{if } \hat{\beta}_j > 0; \\
  \lambda = -n^{-1}X_j^T(Y - X\hat{\beta}) & \text{if } \hat{\beta}_j < 0; \\
  \lambda \geq |n^{-1}X_j^T(Y - X\hat{\beta})| & \text{if } \hat{\beta}_j = 0.
\end{cases}
\]

(b) If \( \lambda > \|n^{-1}X^TY\|_\infty \), prove that \( \hat{\beta}_\lambda = 0 \), where \( \hat{\beta}_\lambda \) is the minimizer of the Lasso problem with regularization parameter \( \lambda \).

(c) If \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) are both minimizers of the Lasso problem, show that they have the same prediction, i.e., \( X\hat{\beta}_1 = X\hat{\beta}_2 \).

**Hint:** Consider the vector \( \tilde{\beta}_\alpha = \alpha\hat{\beta}_1 + (1 - \alpha)\hat{\beta}_2 \) for \( \alpha \in (0, 1) \). Show that \( Q(\tilde{\beta}_\alpha) \) does not depend on \( \alpha \) by using the convexity, where \( Q(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \). The same convexity argument entails that the loss \( L(\alpha) = \frac{1}{2n}\|Y - X\beta_\alpha\|_2^2 \) does not depend on \( \alpha \) and conclude the result from here.

2. Risk properties of Lasso.

Let \( R_n(\beta) = \|Y - X\beta\|_2^2/n \) and \( R(\beta) = ER_n(\beta) \) be the empirical and theoretical risks, and \( \hat{\beta} = \text{argmin}_{\|\beta\|_1 \leq c} R_n(\beta) \) be the Lasso estimator which estimates \( \beta_0 = \text{argmin}_{\|\beta\|_1 \leq c} R(\beta) \).

(a) Consider the in-sample risk \( R_n(\hat{\beta}) \) as an estimator of optimal risk \( R(\beta_0) \). Show that
\[
|R(\beta_0) - R_n(\hat{\beta})| \leq \max_{\|\beta\|_1 \leq c} |R(\beta) - R_n(\beta)| \leq (1 + c)^2 \|\Sigma^* - S^*_n\|_{\max},
\]
where \( Z = \begin{pmatrix} Y \\ X \end{pmatrix}, \Sigma^* = E(ZZ^T) \) and \( S^*_n = n^{-1} \sum_{i=1}^n Z_iZ_i^T \).

**Hint:** Deal with two sides of the inequality separately. For example, \( R(\beta_0) - R_n(\hat{\beta}) = R(\beta_0) - R_n(\beta_0) + R_n(\beta_0) - R_n(\hat{\beta}) \geq R(\beta_0) - R_n(\beta_0) \).

(b) Suppose that \( \|X\|_\infty \leq b \) and \( |Y| \leq b \) (bounded random variables). Use Hoeffding’s inequality to show \( \|\Sigma^* - S^*_n\|_{\max} = O_p(\sqrt{\frac{\log p}{n}}) \).

(c) Consider the lasso of form \( \hat{\beta} = \text{argmin}_{\beta} \{ \frac{1}{2} R_n(\beta) + \lambda \|\beta\|_1 \} \).

If \( \lambda \geq \|n^{-1}X^T(Y - X\beta_0)\|_\infty \), under the restricted eigenvalue condition
\[
\min_{\|\Delta s_0\|_2 \geq \|\Delta s_0\|_1} n^{-1}\|X\Delta\|_2^2/\|\Delta\|_2^2 \geq a,
\]
show that with \( \hat{\Delta} = \hat{\beta} - \beta_0 \) and \( s = |\text{Supp}(\beta_0)| \),
\[
\|\hat{\Delta}\|_2 \leq 8a^{-1}\sqrt{s}\lambda \quad \text{and} \quad \|\hat{\Delta}\|_1 \leq 32a^{-1}s\lambda.
\]
3. Concentration inequalities.

(a) The random vector \( \varepsilon \in \mathbb{R}^n \) is called \( \sigma \)-sub-Gaussian if \( E \exp(a^T \varepsilon) \leq \exp(\|a\|^2 \sigma^2/2), \forall a \in \mathbb{R}^n \). Show that \( E \varepsilon = 0 \) and \( \text{var}(\varepsilon) \leq \sigma^2 I_n \).

**Hint:** Expand exponential functions as infinite series (actually, you only need the condition for \( a \) in a small neighborhood around 0)

(b) Suppose that the random vector \( \mathbf{X} - E \mathbf{X} \) is \( \sigma \)-sub-Gaussian and \( S_n = 1^T \mathbf{X} = \sum_{i=1}^n X_i \). Show that

\[
P\left(n^{-1/2}\|S_n - ES_n\| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right), \quad t > 0.
\]

**Hint:** Use Chebyshev’s inequality

\[
P\left(n^{-1/2}(S_n - ES_n) \geq t\right) \leq \exp(-xt)E\exp\left(xn^{-1/2}(S_n - ES_n)\right)
\]

and optimize the choice of \( x \) after using the moment generating function of sub-Gaussian distributions.

(c) For \( \mathbf{X} \in \mathbb{R}^{n \times p} \) with the \( j \)-th column denoted by \( \mathbf{X}_j \in \mathbb{R}^n \), suppose that \( \|\mathbf{X}_j\|_2^2 = n \) for all \( j \), and \( \varepsilon \in \mathbb{R}^n \) is a \( \sigma \)-sub-Gaussian random vector. Show that there exists a constant \( C > 0 \) such that

\[
P\left(\|n^{-1}\mathbf{X}^T\varepsilon\|_\infty > \sqrt{2(1+\delta)}\sigma\sqrt{\frac{\log p}{n}}\right) \leq Cp^{-\delta}, \quad \forall \delta > 0.
\]

**Hint:** Using the same argument as part (b), we can obtain \( P\{\|\mathbf{b}^T\varepsilon\| \geq t\} \leq 2 \exp\left(-\frac{t^2}{2\sigma^2\|\mathbf{b}\|}\right) \). You can use this without proof.

4. This problem intends to show that the gradient decent method for a convex function \( f(\cdot) \) is a member of majorization-minimization algorithms and has a sublinear rate of convergence in terms of function values. From now on, the function \( f(\cdot) \) is convex and let \( \mathbf{x}^* \in \text{argmin} f(\mathbf{x}) \). Here we implicitly assume the minimum can be attained at some point \( \mathbf{x}^* \in \mathbb{R}^p \).

(a) Suppose that \( f''(\mathbf{x}) \leq L \mathbf{I}_p \) and \( \delta \leq 1/L \). Show that the quadratic function \( g(\mathbf{x}) = f(\mathbf{x}_{i-1}) + f'(\mathbf{x}_{i-1})^T(\mathbf{x} - \mathbf{x}_{i-1}) + \frac{1}{2\delta}\|\mathbf{x} - \mathbf{x}_{i-1}\|^2 \) is a majorization of \( f(\mathbf{x}) \) at point \( \mathbf{x}_{i-1} \), i.e., \( g(\mathbf{x}) \geq f(\mathbf{x}) \) for all \( \mathbf{x} \) and also \( g(\mathbf{x}_{i-1}) = f(\mathbf{x}_{i-1}) \).

(b) Show that gradient step \( \mathbf{x}_i = \mathbf{x}_{i-1} - \delta f'(\mathbf{x}_{i-1}) \) is the minimizer of the majorized quadratic function \( g(\mathbf{x}) \) and hence the gradient descend method can be regarded as a member of MM-algorithms. Use (a) to show that

\[
f(\mathbf{x}_i) \leq g(\mathbf{x}_i) = f(\mathbf{x}_{i-1}) - \frac{1}{2\delta}\|\mathbf{x}_i - \mathbf{x}_{i-1}\|^2.
\]

(c) Show that

\[
f(\mathbf{x}_i) \leq f(\mathbf{x}^*) + \frac{1}{2\delta}\left(\|\mathbf{x}_{i-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}^* - \mathbf{x}_{i-1}\|^2\right).
\]

**Hint:** By convexity, \( f(\mathbf{x}^*) \geq f(\mathbf{x}_{i-1}) + f'(\mathbf{x}_{i-1})^T(\mathbf{x}^* - \mathbf{x}_{i-1}) \) and substitute this into the second part of part (b).
(d) Conclude using (c) that $f(x_k) - f(x^*) \leq \|x_0 - x^*\|^2/(2k\delta)$, namely gradient descent converges at a sublinear rate. (Note: The gradient descent method converges linearly if $f(\cdot)$ is strongly convex.)

5. Let us consider the Zillow data again. We drop the first 3 columns (“(empty)”, “id”, “date”)
and treat “zipcode” as a factor variable. Now, consider the variables

(a) “bedrooms”, “bathrooms”, “sqft living”, and “sqft lot” and their interactions and the
remaining 14 variables in the data, including “zipcode”. (We can use `model.matrix` to
expand factors into a set of dummy variables.)

(b) Add the following additional variables to (a): $X_{12} = I(\text{view} == 0)$, $X_{13} = L^2$, $X_{13+i} = (L - \tau_i)^2$, $i = 1, \cdots, 9$, where $\tau_i$ is $10 \times i^{th}$ percentile and $L$ is the size of living area (“sqft living”).

Compute and compare out-of-sample $R^2$ using ridge regression, Lasso (using R package `glmnet`) and SCAD (using R package `ncvreg`) with regularization parameter chosen by 10 fold cross-validation. Set a random seed by `set.seed(525)` before executing Lasso.