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) 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 \( \alpha\hat{\beta}_1 + (1 - \alpha)\hat{\beta}_2 \) for \( \alpha \in (0, 1) \).

(b) Let \( \hat{\beta} \) be a minimizer of the Lasso problem with \( j \)-th component \( \hat{\beta}_j \). Denote \( X_j \) to be the \( j \)-th column of \( X \). Show that

\[
\begin{align*}
\lambda &= n^{-1}X_j^T(Y - X\hat{\beta}) \quad \text{if } \hat{\beta}_j > 0; \\
\lambda &= -n^{-1}X_j^T(Y - X\hat{\beta}) \quad \text{if } \hat{\beta}_j < 0; \\
\lambda &\geq n^{-1}X_j^T(Y - X\hat{\beta}) \quad \text{if } \hat{\beta}_j = 0.
\end{align*}
\]

(c) 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 \( \lambda \).

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 \( \hat{\beta} = \arg\min_{\| \beta \|_1 \leq c} R_n(\beta) \) be the Lasso estimator which estimates \( \beta_0 = \arg\min_{\| \beta \|_1 \leq c} R(\beta) \).

(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 \( S_n^* = n^{-1} \sum_{i=1}^n Z_i Z_i^T \).

**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 \( \| \Sigma^* - S_n^* \|_{\max} = O_p(\sqrt{\log p / n}) \).

(c) Consider the lasso of form \( \hat{\beta} = \arg\min \left\{ \frac{1}{2} R_n(\beta) + \lambda \| \beta \|_1 \right\} \).

If \( \lambda > \| n^{-1}X^T(Y - X\beta_0) \|_\infty \), under the restricted eigenvalue condition

\[
\min_{\| \Delta S_0 \|_1 \geq \| \Delta S_0 \|_1} n^{-1} \| X\Delta \|_2^2 / \| \Delta \|_2^2 \geq a,
\]

show that with \( \Delta = \hat{\beta} - \beta_0 \) and \( s = \| \text{Supp}(\beta_0) \|, \)

\[
\| \Delta \|_2 \leq 8a^{-1} \sqrt{s\lambda} \quad \text{and} \quad \| \Delta \|_1 \leq 32a^{-1} s\lambda.
\]

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 \( X - EX \) is \( \sigma \)-sub-Gaussian and \( S_n = 1^T X = \sum_{i=1}^n X_i \). Show that
\[
P(\sqrt{n}|S_n - ES_n| \geq t) \leq \exp(- \frac{t^2}{2\sigma^2}), \quad t > 0.
\]
**Hint:** Use Chebyshev’s inequality \( P(\sqrt{n}|S_n - ES_n| \geq t) \leq \exp(-xt)E \exp(x\sqrt{n}|S_n - ES_n|) \) and optimize the choice of \( x \) after using the moment generating function of sub-Gaussian distributions.

(c) For \( X \in \mathbb{R}^{n \times p} \) with the \( j \)-th column denoted by \( X_j \in \mathbb{R}^n \), suppose that \( \|X_j\|^2 = n \) 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
\[
P\left( \|n^{-1} X^T \varepsilon\|_\infty > \sqrt{2(1 + \delta)}\sigma \sqrt{\frac{\log p}{n}} \right) \leq Cp^{-\delta}, \quad \forall \delta > 0.
\]

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 \( x^* \in \text{argmin} f(x) \). Here we implicitly assume the minimum can be attained at some point \( x^* \in \mathbb{R}^p \).

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

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

(c) Use (a) and the convexity of \( f(\cdot) \) to show that
\[
f(x_i) \leq f(x^*) + \frac{1}{2\delta}(\|x_{i-1} - x^*\|^2 - \|x^* - x_i\|^2).
\]

(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 (b): \( X_{12} = I(\text{view} == 0) \), \( X_{13} = L^2 \), \( X_{13+i} = (L - \tau_i)^2 \), \( i = 1, \ldots, 9 \), where \( \tau_i \) is 10 * \( i \)-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 \texttt{glmnet}) and SCAD (using R package \texttt{ncvreg}) with regularization parameter chosen by 10 fold cross-validation. Set a random seed by \texttt{set.seed(525)} before executing Lasso.