STT 873 HW3 (Solution Keys)

This HW is due on Oct 30th.

Question 1(email the R code directly to zhang318@stt.msu.edu and xyy@egr.msu.edu): Compare the performance of LS, Best subset, Ridge Regression, LASSO, PCR and PLS on the Prostate Cancer Data. Use the same procedure on data preparation as used in the textbook (Sections 3.2.1 and 3.3.4). The Prostate Cancer Data are available from the book website www-stat.stanford.edu/ElemStatLearn. Report both the R code and the results. (*10 pts for each methods. Total 60 pts.*)

Solution: Least squares, best subset, ridge regression, LASSO, PCR, and PLS were performed on the Prostate Cancer data. The models were fit using 10 fold cross-validation on a training dataset.

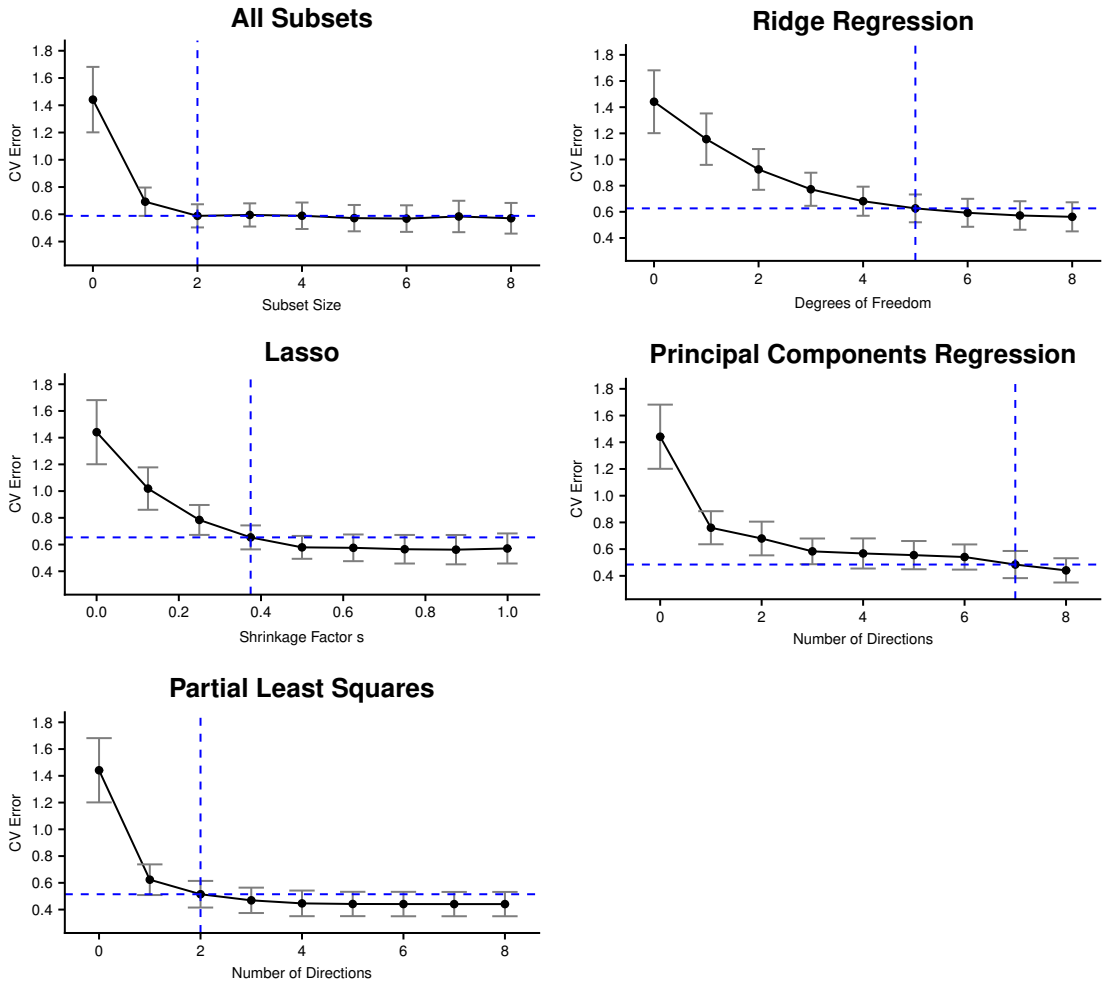


Figure 1: Estimated prediction error curves and their standar errors for the various selection and shrinkage methods.

The estimated prediction error curves are shown in Figure 1 and the estimated coefficients and test error are presented in Table 1. Figure 1 and Table 1 are replicated accurately from Elements of Statistical Learning book (see page 62 for the figure and page 63 for the table).

Table 1: Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.

	LS	BS	Ridge	Lasso	PCR	PLSR
(Intercept)	2.465	2.477	2.464	2.468	2.497	2.467
lcavol	0.680	0.740	0.406	0.533	0.551	0.419
lweight	0.263	0.316	0.226	0.169	0.289	0.345
age	-0.141		-0.043		-0.155	-0.026
lbph	0.210		0.166	0.002	0.214	0.220
svi	0.305		0.230	0.094	0.315	0.243
lcp	-0.288		0.009		-0.062	0.078
gleason	-0.021		0.044		0.228	0.011
pgg45	0.267		0.128		-0.048	0.084
Test Error	0.521	0.492	0.494	0.479	0.529	0.530
Std Error	0.179	0.143	0.162	0.164	0.149	0.152

Question 2 (email the R code directly to zhang318@stt.msu.edu and xyy@egr.msu.edu):

(1) Write a R code to generate the Lasso solution path and analyze the following dataset: https://goo.gl/h779vj (You can compare your result with the R packages 'Lars' or 'elasticnet'. Details of the algorithm can be found in Section 5.6 of the book "Statistical Learning with Sparsity"). (*10 pts.*)

(2) Now let's deal with some highly correlated X. Download data from the following link: https://goo.gl/xSHnWr
find the solution path for Lasso and Elastic net. Report the results and explain the differences. (*10 pts.*)

(3) Report boxplots of 1000 bootstrap realizations of $\hat{\beta}^*(\hat{\lambda}_{\text{CV}})$ similar to Figure 6.4 in the book "Statistical Learning with Sparsity". (*10 pts.*)

(4) For $\lambda = 0.1$, run the post selection inference for Lasso using the second dataset. Report your result using figure similar to Figure 6.12 in the book "Statistical Learning with Sparsity". (*10 pts.*)

Solution:

(1) Figure 2 show the lasso solution path. Figure 2(a) struggles to show the path from $\log(\lambda_k \in [3.5, 2.5]$, and therefore that section of the plot is shown in figure 2(b). This method was compared to the elasticnet and the results are the same.

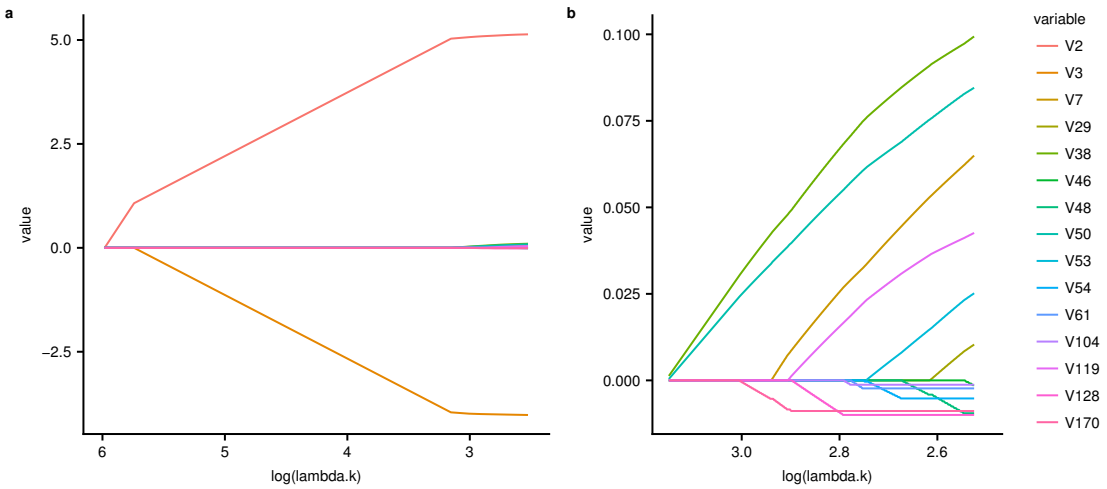


Figure 2: Lasso solution path. (a) full solution path for all variables. (b) zoomed in section of the full solution pat for a more interpretable figure.

(2) Figure 3 shows the Lasso and Elastic net solution paths for highly correlated predictors. To start, note the difference in the scales of the standardized coefficients. Furthermore, Lasso shows some sharp changes in the coefficients, while Elastic net shows smoother changes in the coefficients. Because the data is highly correlated, we see the instability of Lasso. Finally, due to the added term in Elastic net, we see the greater penalty for larger coefficients, leading to relatively larger coefficients for those variables with very little representation in the Lasso.

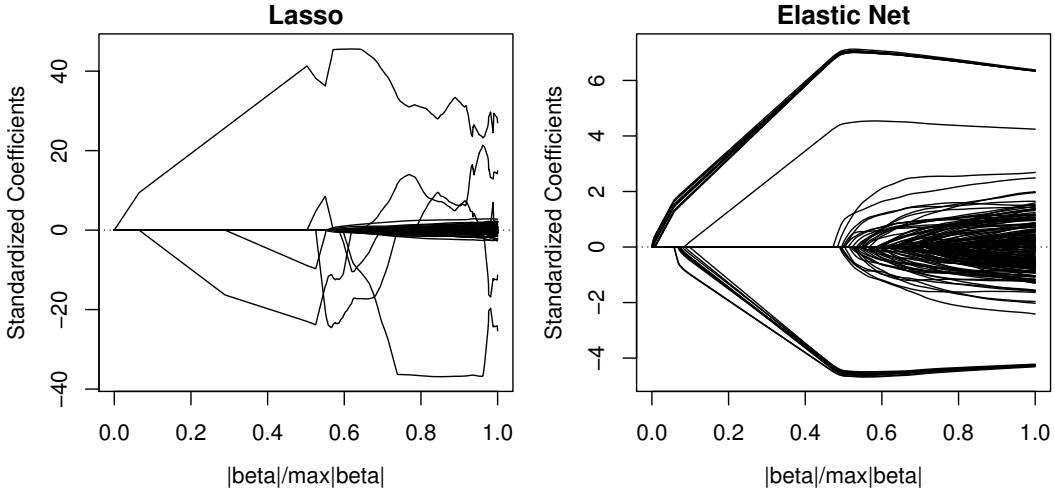


Figure 3: Lasso and Elastic net solution paths for highly correlated predictors

(3) Figure 4 shows boxplots for 1000 relations of $\hat{\beta}^*(\hat{\lambda}_{CV})$ obtained by the nonparametric bootstrap. This corresponds to the re-sampling from the emperical CDF $\hat{F}_N$. Note that only nonzero coefficients were plotted. We can see that many of the coefficients are near zero, but there are a few that appear to be significantly different form zero.

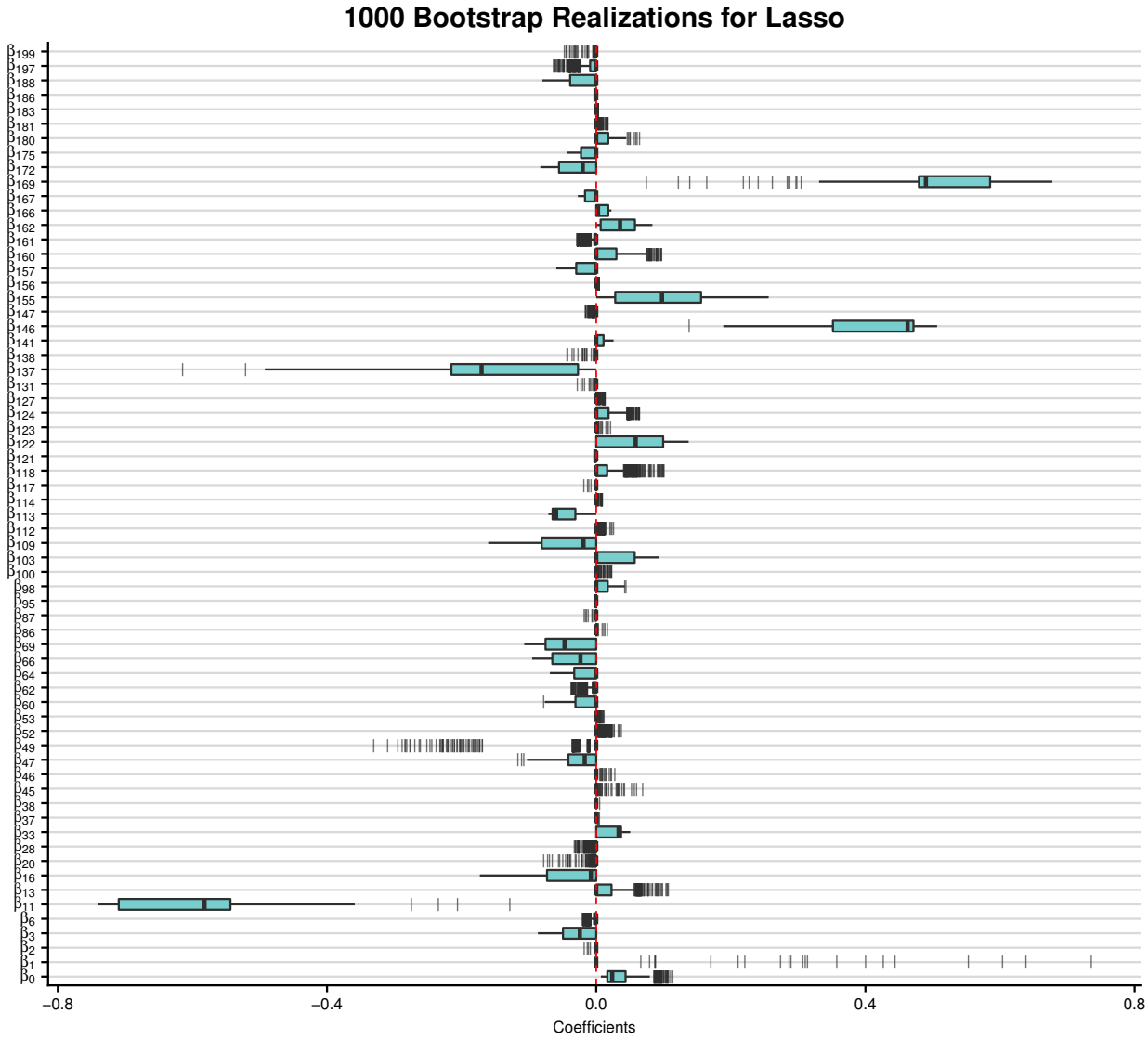
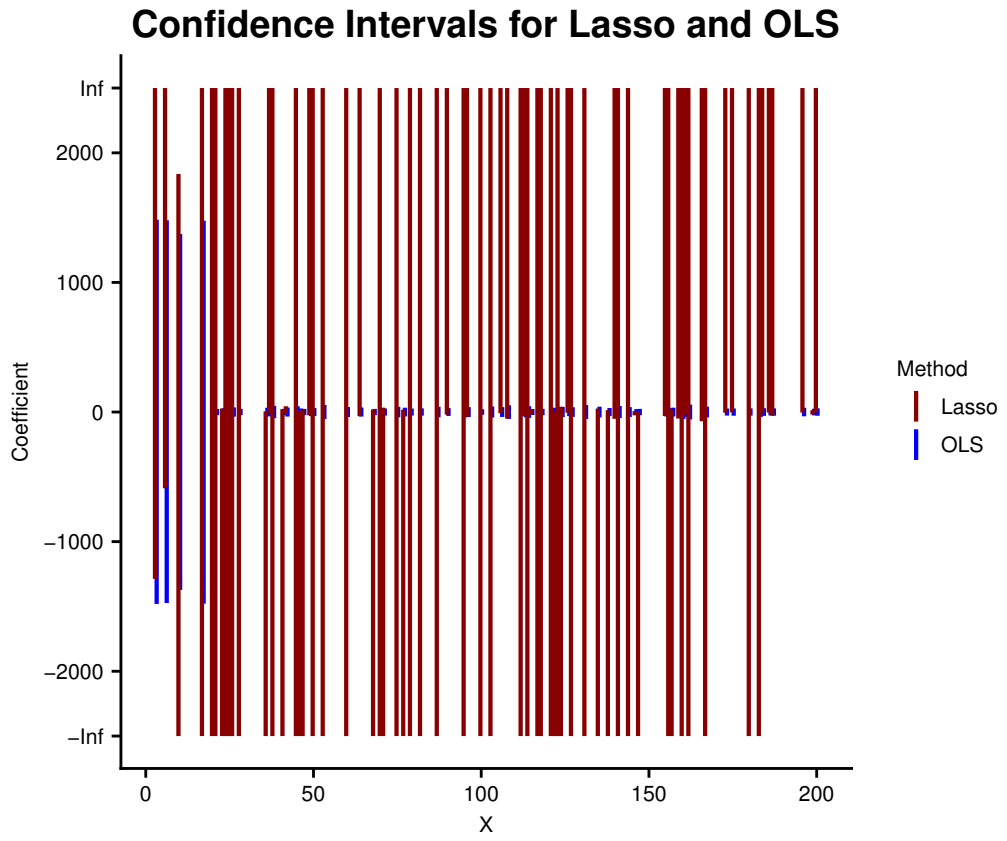


Figure 4: Boxplots of 1000 bootstrap realizations obtained by the nonparametric bootstrap, which corresponds to re-sampling from the empirical CDF.

(4) Figure shows the post selection 95% confidence intervals for Lasso, and they are compared with the 95% confidence intervals form OLS. Note that none of the coefficients from Lasso are significant, and the several of the intervals are unbounded. Moreover, all Lasso intervals are wider that the associated OLS interval. There are suspected to be caused by λ being so small, $\lambda = 0.1$.



Question 3 (Gaussian maximal inequalities): In class, we extensively used the following Gaussian maximal inequality to bound the slow rate, which you will prove here, over the next few parts. If $W_i \sim N(0, \sigma_i^2), i = 1, \ldots, d$ are Gaussian variates, not necessarily independent, then for any $\delta > 0$,

$$P(\max_{1 \le i \le d} |W_i| \le \sigma \sqrt{2 \log(ed/\delta)}) \ge 1 - \delta, \tag{1}$$

where $\sigma = \max_{i=1,\ldots,d} \sigma_i$.

(1) Prove that, for any $t > 0$,

$$P(\max_{1 \le i \le d} |W_i| \ge t) \le 2d \frac{\phi(t/\sigma)}{t/\sigma},$$

where ϕ is the standard normal density. Hint: you may use Mill's inequality.

(2) Using the result from the previous part, plug in $t = \sigma\sqrt{2\log(ed/\delta)}$ and establish (1). (*10 pts.*)

The result (1) is a high-probability bound on the maximum of Gaussians. We can also establish an expectation bound,

$$E(\max_{1\le i\le d}|W_i|) \le \sigma\sqrt{2\log(2d)}. \tag{2}$$

(*10 pts.*)

(3) To prove that, for any $t > 0$

$$E(\max_{1\le i\le d}|W_i|) \le \frac{\log(2d)}{t} + t\sigma^2/2.$$

Hint: use Jensens inequality to argue that

$$\exp(tE(\max_{1\le i\le d}|W_i|)) \le E(\exp(\max_{1\le i\le d} t|W_i|));$$

also, it will help to recall that the moment-generating function of a standard Gaussian variate. Then, you can bound the expectation for the absolute value. (*10 pts.*)

(4) Using the result from the previous part, plug in an appropriate value of t and establish (2). (*10 pts.*)

Solution:

(1)

$$\begin{aligned}
P(\max_{1\le i\le d}|W_i| \ge t) &= P(\cup_{i=1}^{d}|W_i| \ge t) \\
&\le \sum_{i=1}^{d} P(|W_i| \ge t) \\
&= \sum_{i=1}^{d} P(|Z_i| \ge \frac{t}{\sigma_i}) \\
&= dP(|Z| \ge \frac{t}{\sigma}) \quad \text{where } \sigma = \max_{1\le i\le d}\sigma_i \\
&\le d\sqrt{\frac{2}{\pi}}\frac{e^{-\frac{(t/\sigma)^2}{2}}}{t/\sigma} \quad \text{by Mill's inequality} \\
&= 2d\frac{\phi(t/\sigma)}{t/\sigma}
\end{aligned}$$

(2) Plug in $t = \sigma\sqrt{2\log(ed/\delta)}$, then

$$\begin{aligned}
2d\frac{\phi(t/\sigma)}{t/\sigma} &= 2d\frac{\frac{1}{\sqrt{2\pi}}e^{-\log(ed/\delta)}}{\sqrt{2\log(ed/\delta)}} \\
&= \frac{\delta}{e\sqrt{\pi\log(ed/\delta)}} \\
&\le \delta
\end{aligned}$$

By (1)

$$\begin{aligned} P(\max_{1\le i\le d}|W_i| \le \sigma\sqrt{2\log(ed/\delta)}) &= 1 - P(\max_{1\le i\le d}|W_i| \ge \sigma\sqrt{2\log(ed/\delta)}) \\ &\ge 1-\delta \end{aligned}$$

(3) By Jensens inequality, we have

$$\begin{aligned} \exp(tE(\max_{1\le i\le d}|W_i|)) &\le E(\exp\{\max_{1\le i\le n} t|W_i|\}) \\ &\le \sum_{i=1}^{d} E(e^{t|W_i|}) \\ &= \sum_{i=1}^{d} 2\int_0^\infty e^{tx}\frac{1}{\sqrt{2\pi\sigma_i^2}}e^{-\frac{x^2}{2\sigma_i^2}}dx \\ &= \sum_{i=1}^{d} 2e^{\frac{\sigma_i^2 t^2}{2}}\int_0^\infty \frac{1}{\sqrt{2\pi\sigma_i^2}}e^{-\frac{(x-\sigma_i^2 t)^2}{2\sigma_i^2}}dx \\ &\le 2de^{\frac{\sigma^2 t^2}{2}} \end{aligned}$$

Hence,

$$E(\max_{1\le i\le d}|W_i|) \le \frac{\log(2d)}{t} + t\sigma^2/2$$

(4) If we put $t = \sqrt{2\log(2d)}/\sigma$, then

$$\begin{aligned} E(\max_{1\le i\le d}|W_i|) &\le \frac{\log(2d)}{\sqrt{2\log(2d)}/\sigma} + \frac{\sqrt{2\log(2d)}}{\sigma}\frac{\sigma^2}{2} \\ &= \sigma\sqrt{2\log(2d)} \end{aligned}$$

Question 4 (In-sample risk for the lasso): Assume that $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, n$ are i.i.d. pairs satisfying $y_i = x_i^T\beta_0 + \epsilon_i$, where $\beta_0 \in \mathbb{R}^d$ is the unknown parameter. We also assume $x_i \sim P_X$ and $\epsilon_i \sim N(0, \sigma^2)$ with the predictors and errors being independent. Additionally, assume that P_X is a distribution supported on $[-M, M]^d$. Let $\hat{\lambda}$ be the lasso estimator in constrained form,

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^d}{\operatorname{argmin}} \|y - X\beta\|_2^2 \text{ subject to } \|\beta\|_1 \le t,$$

where $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ is the response vector and $X \in \mathbb{R}^{n\times d}$ is the matrix of predictors, with rows $x_i, i = 1, \ldots, n$.
Prove that the lasso estimator, with $t = \|\beta_0\|_1$, has in-sample risk satisfying

$$\frac{1}{n}\mathbb{E}\|X\hat{\beta} - X\beta_0\|_2^2 \le M\sigma\|\beta_0\|_1\sqrt{\frac{2\log(2d)}{n}},$$

where the expectation is taken over the training data $(x_i, y_i), i = 1, \ldots, n$. Hint: follow the same strategy we used in class to derive the slow rate for the lasso estimator in bound form. Take an expectation where appropriate (rather than invoking high-probability arguments as we did in class), and apply the result in (2).

Solution:

$$\begin{aligned}
& \|y - X\hat{\beta}\|_2^2 \leq \|y - X\beta_0\|_2^2 \\
\Rightarrow \quad & \|y\|_2^2 + \|X\hat{\beta}\|_2^2 - 2y^T X\hat{\beta} \leq \|y\|_2^2 + \|X\beta_0\|_2^2 - 2y^T X\beta_0 \\
\Rightarrow \quad & \|X\hat{\beta}\|_2^2 - \|X\beta_0\|_2^2 \leq 2y^T(X\hat{\beta} - X\beta_0) = 2(X\beta_0 + \epsilon)^T(X\hat{\beta} - X\beta_0) \\
\Rightarrow \quad & \|X\hat{\beta}\|_2^2 - 2(X\beta_0)^T(X\hat{\beta}) + \|X\beta_0\|_2^2 \leq 2\epsilon^T(X\hat{\beta} - X\beta_0) \\
\Rightarrow \quad & \|X\hat{\beta} - X\beta_0\|_2^2 \leq 2(X^T\epsilon)^T(\hat{\beta} - \beta_0)
\end{aligned}$$

By Holder's inequality, we have

$$\begin{aligned}
\|X\hat{\beta} - X\beta_0\|_2^2 &\leq 2\|X^T\epsilon\|_\infty \|\hat{\beta} - \beta_0\|_1 \\
&\leq 2\|X^T\epsilon\|_\infty (\|\hat{\beta}\|_1 + \|\beta_0\|_1) \\
&\leq 4\|X^T\epsilon\|_\infty \|\beta_0\|_1 \quad \text{since} \quad \|\beta\|_1 \leq t = \|\beta_0\|_1 \\
&\leq 4 \max_{1 \leq j \leq d} |(\sum_{i=1}^n x_i\epsilon_i)_j| \|\beta_0\|_1
\end{aligned}$$

Now, dividing n and taking expectation on both side,

$$\frac{1}{n} E\|X\hat{\beta} - X\beta_0\|_2^2 \leq \frac{4}{n}\|\beta_0\|_1 E\left(\max_{1 \leq j \leq d} |(\sum_{i=1}^n x_i\epsilon_i)_j|\right) \tag{3}$$

Note that, by conditional expectation we have

$$E\left(\max_{1 \leq j \leq d} |(\sum_{i=1}^n x_i\epsilon_i)_j|\right) = E\left(E_{P_X}\left(\max_{1 \leq j \leq d} |(\sum_{i=1}^n x_i\epsilon_i)_j|\right)\right). \tag{4}$$

Given $x_i \sim P_X$, let $W = \sum_{i=1}^n x_i\epsilon_i$, then $W|x_i \sim N(0, \Sigma)$, where $\Sigma = \sigma^2 \sum_{i=1}^n x_i x_i^T$. Moreover, $W_j = (\sum_{i=1}^n x_i\epsilon_i)_j$ and $W_j|x_i \sim N(0, \Sigma_{(j,j)})$, where $\Sigma_{(j,j)} = \sigma^2 \sum_{i=1}^n x_{ij}^2 \leq n\sigma^2 M^2$. Use the result in question (3), we have

$$\begin{aligned}
E_{P_X}\left(\max_{1 \leq j \leq d} |(\sum_{i=1}^n x_i\epsilon_i)_j|\right) &= E_{P_X}\left(\max_{1 \leq j \leq d} |W_j|\right) \\
&\leq \sqrt{\Sigma_{(j,j)}}\sqrt{2\log(2d)} \\
&\leq \sqrt{n}\sigma M\sqrt{2\log(2d)}.
\end{aligned}$$

Using above inequality and equation (4), plug into equation (3), we establish that

$$\frac{1}{n} E\|X\hat{\beta} - X\beta_0\|_2^2 \leq 4M\sigma\|\beta_0\|_1\sqrt{\frac{2\log(2d)}{n}}$$