Consider the sparse linear regression model presented in the lecture notes. Let

$$y = X\beta^* + w, \tag{1}$$

for a design matrix $X \in \mathbb{R}^{n\times d}$, a Gaussian vector $w \sim N(0, Id_n)$ and a k–sparse vector β^* (for some $1 \le k \le d$) such that $||\beta^*||^2 \le 10k$ and for any $i \in \text{supp}\{\beta^*\}$,$|\beta_i^*| \ge 1$. Given the pair (y, X) the goal is to find $\hat{\beta} \in \mathbb{R}^d$ such that $X\hat{\beta}$ is close to $X\beta^*$.

The goal of the exercise is to study the guarantees of the following nave algorithm, which iteratively applies least square regression to each coordinate. Here, for $i \in [d]$, we denote by X_i the columns of X.

Assignment 1 (10 points)

Consider the settings of Theorem 1. Let $i \in \text{supp}\{\beta^*\}$. Show that if $||X_i|| \ge 10\sqrt{\ln d}$, then $|\hat{\beta}_i| \ge 1/2$ with probability at least $1 - O(d^{-10})$.

Suppose $|\hat{\beta}_i| < 1/2$. Assuming $||X_i|| \ge 10\sqrt{\ln d}$ from the Algorithm 1. step 2 it follows that $|s_i| < 1/2$ (because otherwise $|\hat{\beta}_i| = |s_i| \ge 1/2$. So now it's equivalent to prove that

$$P \equiv \mathbb{P}(|s_i| < 1/2 \big| ||X_i|| \ge 10\sqrt{\ln d}) \le O(d^{-10}).$$

Notice, that $X_i^\top X_i$ is scalar and equals to $||X_i||^2$, therefore we need probability of

$$|s_i| = |(X_i^\top X_i)^{-1} X_i^\top (X\beta^* + w)| = \frac{|X_i^\top (X\beta^* + w)|}{||X_i||^2} \le 1/2.$$

From the formulation of the Theorem 1. we have $\langle X_i, X_j \rangle = 0 \; \forall i \ne j$, and because $X\beta^* = \sum_{i=1}^d \beta_i^* \cdot X_i$, therefore we have

$$|s_i| = \left| \frac{X_i^\top w}{||X_i||^2} + \beta_i^* \right| \le 1/2. \tag{2}$$

Because from formulation of the Theorem 1 we have that for $i \in \text{supp}\{\beta^*\}$ (from Assignment 1. formulation) $|\beta_i^*| \ge 1$ and that from triangle inequality

$$\left| \frac{X_i^\top w}{||X_i||^2} \right| \ge 1/2,$$

and its probability $P_1 \equiv \mathbb{P}\left(\left| \frac{X_i^\top w}{||X_i||^2} \right| \ge 1/2 \right) \ge P$, as it follows from the needed statement. Now we are estimating P_1.

Dividing X_i by $||X_i||$ we have $u = X_i/||X_i||$ - unit vector independent from w. Therefore, using the Fact. 3 (where $\underline{g}$ is $w \sim N(0, 1^2 \cdot Id_n)$), we have $\langle u, w \rangle \sim N(0, 1)$ ($X_i^\top w$ is a scalar).

As $||X_i|| \ge 10\sqrt{\ln d}$, we have that

$$P_1 \equiv \mathbb{P}(|\langle u, w \rangle| \ge 1/2 \cdot 10\sqrt{\ln d}) \ge P$$

Using after that the Fact 1., inserting as z there our $\langle u, w \rangle$, we have that ($t = 5\sqrt{\ln d}$ and $\sigma = 1$)

$$P_1 \equiv \mathbb{P}(|\langle u, w \rangle| \ge 5\sqrt{\ln d}) \le 2\exp(-1/2 \cdot (5\sqrt{\ln d})^2) = 2d^{-12.5} = O(d^{-10})$$

Because we went from P to P_1 only expanding the set of interest, and got $P \le P_1 = O(d^{-10})$, therefore $P = O(d^{-10})$ too, what needed.

Assignment 2 (10 points)

Consider the settings of Theorem 1. Let $i \in \operatorname{supp}\{\beta^*\}$. Show that

$$||X_i(\beta_i^* - \hat{\beta}_i)||^2 \leq O(|\beta_i^*|^2 \ln d)$$

with probability at least $1 - O(d^{-10})$.

a) Suppose first that $||X_i|| < 10\sqrt{\ln d}$. From Algorithm 1 then it follows that $\hat{\beta}_i = 0$ and therefore $||X_i(\beta_i^* - \hat{\beta}_i)||^2 = ||X_i\beta_i^*||^2 \leq ||X_i||^2|\beta_i^*|^2 \leq |\beta_i^*|^2 100 \ln d = O(|\beta_i^*|^2 \ln d)$.

b) The other case, $||X_i|| \geq 10\sqrt{\ln d}$. Then from the Assignment 1 with probability $1 - O(d^{-10})$ we have $\hat{\beta}_i \geq 1/2$ and therefore, as it was obtained at the step 1, analogous to (2)

$$\hat{\beta}_i = \frac{X_i^\top(X_i\beta_i^* + w)}{||X_|||^2} = \beta_i^* + \frac{X_i^\top w}{||X_i||^2} \Rightarrow (\beta_i^* - \hat{\beta}_i) = -\frac{X_i^\top w}{||X_i||^2}.$$

And

$$||X_i(\beta_i^* - \hat{\beta}_i)||^2 = \left|\left|\frac{X_i X_i^\top w}{||X_i||^2}\right|\right|^2.$$

Dividing every X_i in the nominator by its norm, with $u_i = \frac{X_i}{||X_i||}$ — unit vector, we have to show that

$$||u_i u_i^\top w||^2 \leq O(|\beta_i^*|^2 \ln d)$$

From the Fact 3. using w as g we have, that $u_i^\top w \sim N(0, 1)$. Therefore as $||u_i|| = 1$ we have to show

$$||g||^2 \leq O(|\beta_i^*|^2 \ln d),\ g \sim N(0, 1).$$

But g is scalar, using the Fact 1 we have that $|g| \geq t,\ t \geq 0$ with probability not more than $2\exp(-t^2/2)$. Inserting there e.g. $t = 10\sqrt{\ln d}$, we have with probability not more than $2d^{-50}$ that $|g|^2$ is more than $100 \ln d$, which is $O(|\beta_i^*|^2 \ln d)$.

Therefore, in this case from Assignment 1 we had that if $||X_i|| \geq 10\sqrt{\ln d}$, then with probability at least $1 - 2d^{-12.5}$ we will have $\hat{\beta}_i \neq 0$. But in this case, as we've shown, still the needed inequality holds with probability $1 - 2d^{-50}$, therefore in this case full probability of this is $(1 - 2d^{-12.5})(1 - 2d^{-50}) = 1 - 2d^{-12.5} - 2d^{-50} + 4d^{-62.5} = 1 - O(d^{-10})$.

As in the case a) it holds with probability 1, and here it's at least $1 - O(d^{-10})$, then overall probability is not less than $1 - O(d^{-10})$ independently of probabilities of case a) and b) (which don't intersect and give the whole set of events) (formal proof of obvious fact: $\mathbb{P}(\text{ineq. holds}) = \mathbb{P}(\text{ineq. holds and a})) \cdot \mathbb{P}(a) + \mathbb{P}(\text{ineq. holds and b})) \cdot \mathbb{P}(b) \geq \max\{\mathbb{P}(\text{ineq. holds and a})) \cdot \mathbb{P}(a), \mathbb{P}(\text{ineq. holds and b})) \cdot \mathbb{P}(b)\} \geq \{\text{because } \mathbb{P}(a) + \mathbb{P}(b) = 1; P(a), P(b) \geq 0\} \geq \min\{1 - O(d^{-10}), 1\} = 1 - O(d^{-10})$).

Assignment 3 (5 points)

Consider the settings of Theorem 1. Let $j \in [d] \setminus \operatorname{supp}\{\beta^*\}$. Show that $\hat{\beta}_j = 0$ with probability at least $1 - O(d^{-10})$.

In the Assignment 1 we've shown that

$$\mathbb{P}\left(\frac{|X_i^\top w|}{||X_i||^2} \geq 1/2\right) = O(d^{-10})$$

without using that $i \in \operatorname{supp}\{\beta_i^*\}$. Therefore

$$\mathbb{P}\left(\frac{|X_i^\top w|}{||X_i||^2} < 1/2\right) = 1 - O(d^{-10}).$$

But for $i \notin \mathrm{supp}\{\beta_i^*\}$ we have that $\beta_i^* = 0$ and therefore from (2)

$$|s_i| = \left| \frac{X_i^\top w}{||X_i||^2} + \beta_i^* \right| = \frac{|X_i^\top w|}{||X_i||^2}.$$

And probability of $|s_i| < 1/2$, which from the step 2 of the Algorithm 1 is not less that needed probability, is therefore at least $1 - O(d^{-10})$.

Assignment 4 (10 points)

Use Assignments 2 and 3 to prove Theorem 1

Using the fact that the multiplication of matrix and vector is the linear combination of columns with corresponding coefficient from vector elements, and also using orthogonality of columns (from the Theorem 1. formulation), we have

$$\Delta \equiv \frac{1}{n}||X(\beta^* - \hat{\beta})||^2 = \frac{1}{n}||\sum_{i=1}^{d} X_i(\beta_i^* - \hat{\beta}_i)||^2 = \frac{1}{n}\sum_{i=1}^{d} ||X_i(\beta_i^* - \hat{\beta}_i)||^2 + \frac{1}{n}\sum_{i,j=1;i\neq j}^{d} \underbrace{\langle X_i, X_j\rangle}_{0}(\beta_i^* - \hat{\beta}_i)(\beta_j^* - \hat{\beta}_j)$$

i.e.

$$\Delta = \frac{1}{n}\sum_{i=1}^{d} ||X_i(\beta_i^* - \hat{\beta}_i)||^2 \leq_?^? O(\frac{k}{n}\ln d)$$

Denoting $||X_i(\beta_i^* - \hat{\beta}_i)||^2$ as Δ_i and splitting the sum by the fact that the corresp. index is in support $S \equiv \mathrm{supp}\{\beta^*\}$:

$$\Delta = \frac{1}{n}\left(\sum_{i \in S}\Delta_i + \sum_{j \notin S}\Delta_j\right)$$

1) Let's look at the second sum.

   From the Assignment 3. we have that every $\Delta_j, j \notin S$ has in it $\hat{\beta}_j = 0$ w.h.p. But as it's not in support, β_j^* is 0 too, therefore with probability at least $1 - O(d^{-10})$ we have $\Delta_j = 0$. Denote as $P_1^{(j)}$ probability to have non-zero in Δ_i, which from Assignment 3 as complement to "having zero" is $1 - (1 - O(d^{-10})) = O(d^{-10})$.

2) Let's look at the first sum.

   From the Assignment 2. we have the estimate

   $$\Delta_i \leq O(|\beta_i^*|^2 \ln d),\ i \in S \tag{3}$$

   with probability $1 - O(d^{-10})$. Therefore, assuming it's right for all $i \in S$ and summing these, we'll have

   $$\sum_{i \in S}\Delta_i \leq O\left(\sum_{i \in S}|\beta_i^*|^2 \ln d\right) = O\left(\ln d \sum_{i \in S}|\beta_i^*|^2\right)$$

   Because $\beta_i^* = 0, \forall i \notin S$ and from the Theorem 1. formulation $||\beta^*||^2 \leq 10k$, we have

   $$\sum_{i \in S}\Delta_i = O(\ln d||\beta^*||^2) \leq O(\ln d \cdot 10k) = O(k \ln d)$$

Therefore,

$$\frac{1}{n}\sum_{i\in S}\Delta_i \le O\left(\frac{k}{n}\ln d\right).$$

Denote $P_2^{(i)}$ as probability of Δ_i not complying with (3) for some $i \in S$. It is from Assignment 2 equals (as complement) to $1-(1-O(d^{-10})) = O(d^{-10})$.

Denote now P as probability to have at least one zero in 1) or at least one not complying in 2). It is from the union bound and the fact that $|S| = k$, $|\bar{S}| = d-k$ (of $j \notin S$):

$$P = \bigcup_{j\notin S} P_1^{(j)} \bigcup_{i\in S} P_2^{(i)} \le (d-k)\cdot O(d^{-10}) + k\cdot O(d^{-10}) = O(d^{-9}). \tag{4}$$

So, probability of not having any zero in 1) and having all in 2) complying with (3) is as complement $1-O(d^{-9}) \to 1$ with $d\to\infty$. And this event will mean that $\Delta \le O(\frac{k}{n}\ln d)$ (the first sum of deltas is 0, the second is of this order).

Assignment 5

Consider the settings of Theorem 2. Let $i \in [d]$. Show that, given $||X_i||$, if $||X_i|| \neq 0$,

$$s_i - \beta_i^* \sim \mathcal{N}(0,\sigma^2),$$

where $\sigma^2 = \frac{1}{||X_i||^2}\left(1+\sum_{j\in\text{supp}\{\beta^*\}\setminus\{i\}}(\beta_j^*)^2\right)$.

Note that

$$X_i^\top y \equiv X_i^\top\left(\sum_{j\in[d]} X_j\beta_j^* + w\right) = X_i^\top\left(\sum_{j\in\text{supp}\{\beta^*\}} X_j\beta_j^* + w\right),$$

as $\beta_j^* = 0$ for $j \notin \text{supp}\{\beta^*\}$.

$$s_i - \beta_i^* = \frac{X_i^\top\left(\sum_{j\in\text{supp}\{\beta^*\}} X_j\beta_j^* + w\right)}{||X_i||^2} - \beta_i^* = \frac{X_i^\top\left(\sum_{j\in\text{supp}\{\beta^*\}\setminus\{i\}} X_j\beta_j^*\right)}{||X_i||^2} + \frac{X_i^\top w}{||X_i||^2}.$$

Denoting $g \equiv \frac{X_i}{||X_i||}$ - unit vector and taking as u - independent from g normal vector $\frac{\beta_j^*}{||X_i||}X_j \sim \mathcal{N}\left(0, \frac{(\beta_j^*)^2}{||X_i||^2}\cdot Id_n\right)$ $(j\neq i)$ (quasi–depending on i variance is just a constant, normal distribution X_j stays independent from X_i) and therefore using the Fact 3, we have

$$a_j \equiv \langle g,u\rangle = \frac{X_i^\top}{||X_i||}X_j\frac{\beta_j^*}{||X_i||} \sim \mathcal{N}\left(0,\frac{(\beta_j^*)^2}{||X_i||^2}\right), \qquad j\in\text{supp}\{\beta^*\}\setminus\{i\}.$$

Repeating similar with $u = \frac{1}{||X_i||}w$ independent from g, we have

$$a_0 \equiv \langle g,u\rangle = \frac{X_i^\top w}{||X_i||^2} \sim \mathcal{N}\left(0,\frac{1^2}{||X_i||^2}\right).$$

Adding these two results, we have needed, because all $a_\bullet$ are independent (as all X_j and w are independent of each other) normal distributions with mean 0, therefore the variance of result is a sum of variances, and

$$\sigma^2 = \sum_{j\in\{0\}\cup\text{supp}\{\beta^*\}\setminus\{i\}} a_j = \frac{1}{||X_i||^2}\left(1+\sum_{j\in\text{supp}\{\beta^*\}\setminus\{i\}}(\beta_j^*)^2\right)$$

with $s_i - \beta_i^* \sim \mathcal{N}(0, \sigma^2)$.

As probability of the normal vector to have norm 0 is zero, we assume further in probabilistic reasonings that $||X_i|| \neq 0$.

Assignment 6

Consider the settings of Theorem 2. Let $i \in \operatorname{supp}\{\beta^*\}$. Show that there exists a constant $c > 0$ (not depending on d, n, k or i) such that

$$|\beta_i^* - \hat{\beta}_i| \geq c\sqrt{\frac{k}{n}} \tag{5}$$

with probability at least 0.99.

We know that for

$$i \in \operatorname{supp}\{\beta^*\} \Rightarrow |\beta_i| \geq 1 \tag{6}$$

Also

$$||\beta^*||^2 \leq 10k \tag{7}$$

a) Suppose at Step 2 of the Algorithm 1 we obtained $\hat{\beta}_i = 0$, therefore either $||X_i|| < 10\sqrt{\ln d}$ or $|s_i| < 1/2$ or both. Then, we obtain

$$|\beta_i^* - \hat{\beta}_i| = |\beta_i^* - 0| \geq 1.$$

b) Otherwise, we have

$$\begin{cases} ||X_i|| \geq 10\sqrt{\ln d} \\ |s_i| \geq 1/2 \end{cases}$$

And from this, Assignment 5, (7) and using the condition that $|\operatorname{supp}\{\beta^*\}| = k$ and $|\beta_i^*| \geq 1, i \in \operatorname{supp}\{\beta^*\}$

$$\begin{cases} \hat{\beta}_i - \beta_i^* = s_i - \beta_i^* \sim \mathcal{N}(0, \sigma^2) \\ \sigma^2 = \frac{1}{||X_i||^2}\left(1 + \sum_{j \in \operatorname{supp}\{\beta^*\} \setminus \{i\}} (\beta_j^*)^2\right) \leq \frac{||\beta^*||^2}{||X_i||^2} \leq \frac{10k}{(10\sqrt{\ln d})^2} \Rightarrow \sigma \leq \frac{\sqrt{10k}}{10\sqrt{\ln d}} = \sqrt{\frac{k}{10 \ln d}} \end{cases}$$

From the left inequality of Fact 1 taking $z = \hat{\beta}_i - \beta_i^*$ and $t = ct'$ such that $ct'\sigma = c\sqrt{\frac{k}{n}}$ and c - is a needed constant which we determine later, we have

$$\mathbb{P}\left(|\hat{\beta}_i - \beta_i^*| \geq c\sqrt{\frac{k}{n}}\right) \geq 1 - 2c\frac{\sqrt{\frac{k}{n}}}{\sigma} \geq 1 - 2c\frac{\sqrt{\frac{k}{n}}}{\sqrt{\frac{k}{10 \ln d}}} = 1 - 2c\sqrt{\frac{10 \ln d}{n}} \tag{8}$$

Using conditions $n \geq 1000k \ln d$ and $k \geq 1$ we obtain

$$\mathbb{P}\left(|\hat{\beta}_i - \beta_i^*| \geq c\sqrt{\frac{k}{n}}\right) \geq 1 - 2c\sqrt{\frac{1}{1000k}} = 1 - 2c\frac{1}{\sqrt{1000k}} \geq 1 - 2c\frac{1}{\sqrt{1000}} \tag{9}$$

For probability at least 0.99 we And because additionally $d \geq 1000$, we have

$$c\sqrt{\frac{k}{n}} \leq c\sqrt{\frac{1}{1000 \ln d}} \leq c\sqrt{\frac{1}{1000 \ln 1000}}. \tag{10}$$

So, choosing c such that in (9) we have probability more than 0.99 and such that $c\sqrt{\frac{k}{n}} \leq 1$, we have that in a) we will have needed condition with probability 1, in b it would be at least 0.99, so in general we'll have for corresp. probabilities $\mathbb{P}(a), \mathbb{P}(b) \geq 0$, $\mathbb{P}(a) + \mathbb{P}(b) = 1$ (a and b are mutually excluding and full in union):

$$\mathbb{P}\left(|\hat{\beta}_i - \beta_i^*| \geq c\sqrt{\frac{k}{n}}\right) = \underbrace{\mathbb{P}\left(|\hat{\beta}_i - \beta_i^*| \geq c\sqrt{\frac{k}{n}}\middle| a\right)}_{=1} \mathbb{P}(a) + \underbrace{\mathbb{P}\left(|\hat{\beta}_i - \beta_i^*| \geq c\sqrt{\frac{k}{n}}\middle| b\right)}_{\geq 0.99} \mathbb{P}(b) \geq \min\{1, 0.99\} = 0.99$$

So, we can achieve it by choosing $c = \frac{\sqrt{1000}}{2}$. Therefore $1 - 2c\frac{1}{\sqrt{1000}} = 0.99$ and $c\sqrt{\frac{1}{1000\ln 1000}} = \frac{1}{2\sqrt{\ln 1000}} < \frac{1}{2\sqrt{\ln 3^6}} < \frac{1}{2\sqrt{\ln e^6}} = \frac{1}{2\sqrt{6}} < 1$ obviously.

So, taking independent constant $c = \frac{\sqrt{1000}}{2}$ we achieved the needed estimate.

Assignment 7

Consider the settings of Theorem 2. Show that with probability at least 0.99*,* $\text{supp}\{\hat{\beta}\} \subseteq \text{supp}\{\beta^*\}$.
Let's call $P = \mathbb{P}(\text{supp}\{\hat{\beta}\} \subseteq \text{supp}\{\beta^*\})$.
Note, that it's equal to

$$\mathbb{P}\left(\bigcap_{i\in[d]} \beta_i^* \neq 0, \hat{\beta}_i \neq 0\right),$$

which is as a complement equals to

$$1 - \mathbb{P}\left(\bigcup_{i\in[d]} \beta_i^* = 0 \text{ or } \hat{\beta}_i = 0\right).$$

And this using that $\beta_i^* = 0$ only for $i \notin S$ and from the union bound could be estimated as

$$P = 1 - \mathbb{P}\left(\bigcup_{i\in[d]} \beta_i^* = 0 \text{ or } \hat{\beta}_i \neq 0\right) = 1 - \mathbb{P}\left(\bigcup_{i\notin S} \beta_i^* = 0 \text{ or } \hat{\beta}_i \neq 0\right) \geq 1 - \sum_{i\notin S} \mathbb{P}\left(\beta_i^* = 0 \text{ or } \hat{\beta}_i \neq 0\right) \tag{11}$$

Because $\beta_i^* = 0$ only for $i \notin \text{supp}\{\beta^*\} \equiv S$, therefore

$$P \geq 1 - \sum_{i\notin S} \mathbb{P}\left(\beta_i^* = 0, \hat{\beta}_i \neq 0\right) \overset{?}{\underset{?}{\geq}} 0.99 \tag{12}$$

Using Bayes' rule:

$$\mathbb{P}\left(\beta_i^* = 0, \hat{\beta}_i \neq 0\right) = \mathbb{P}\left(\underbrace{\hat{\beta}_i \neq 0 \middle| \beta_i^* = 0}_{\text{event } A_i}\right) \cdot \underbrace{\mathbb{P}\left(\beta_i^* = 0\right)}_{=1\ \forall i\notin S} \tag{13}$$

Therefore we want to show

$$\sum_{i\notin S} \mathbb{P}(A_i) \leq 0.01, \tag{14}$$

which follows (as $|[d]/S| = d - k$) from proving

$$\mathbb{P}(A_i) \leq \frac{0.01}{d-k}, \forall i \notin S \tag{15}$$

If $d = k$, then in (14) we simply have some over empty set, which is 0, and the inequality holds. Therefore further we consider $k < d$ and look at $i \notin S$.

Observe that $\hat{\beta}_i \neq 0$ for some i, following the Step 2 of the Algorithm 1, means that ($||X_i|| = 0$ with 0 prob.)

$$A_i \leftrightarrow \begin{cases} ||X_i|| \geq 10\sqrt{\ln d} \\ |s_i| \geq 1/2 \\ \hat{\beta}_i = s_i \end{cases}$$

Note that as $A_i \leftrightarrow (|s_i| \geq 1/2$ and $||X_i|| \geq 10\sqrt{\ln d})$, therefore

$$\mathbb{P}(A_i) = \mathbb{P}\left(|s_i| \geq 1/2 \text{ and } ||X_i|| \geq 10\sqrt{\ln d}\right) = \mathbb{P}\left(\underbrace{|s_i| \geq 1/2}_{\text{event } B_i}\middle| ||X_i|| \geq 10\sqrt{\ln d}\right)\mathbb{P}\left(\underbrace{||X_i|| \geq 10\sqrt{\ln d}}_{\text{event } C_i}\right) \tag{16}$$

And

$$\mathbb{P}\left(||X_i|| \geq 10\sqrt{\ln d}\right) = \mathbb{P}\left(||X_i||^2 \geq 100\ln d\right) \tag{17}$$

Observe that

$$100\ln d < \frac{1000 \cdot 1 \cdot \ln d}{2} \leq \frac{1000k\ln d}{2} \leq \frac{n}{2} \tag{18}$$

Using this and the Fact 4 with taking $g = \frac{X_i}{||X_i||} \in \mathbb{R}^n$, which is normal vector from $\mathbb{N}(0, Id_n)$ of length $n \geq 1000k\ln n$ from the theorem description, we have

$$\mathbb{P}(C_i) \equiv \mathbb{P}\left(||X_i||^2 \geq 100\ln d\right) \geq \mathbb{P}\left(||X_i||^2 \geq \frac{n}{2}\right) \geq \mathbb{P}\left(\frac{n}{2} \leq ||X_i||^2 \leq 2n\right) \geq 1 - \exp\left(-\frac{n}{100}\right) \tag{19}$$

Note that because $k \geq 1$ we have $\frac{n}{100} \geq 10k\ln d > 10 \cdot 1 \cdot \ln 1000 > 10 \cdot \ln 3^6 > 10 \cdot \ln e^6 = 60$, and

$$\mathbb{P}(C_i) \geq 1 - \exp(-60) \geq 1 - (2^4)^{-15} \geq 1 - 10^{-15} \tag{20}$$

Now we'll estimate $\mathbb{P}\left(B_i\middle|C_i\right)$. Let's introduce event $D_i \equiv \left(\frac{n}{2} \leq ||X_i||^2 \leq 2n\right)$. We know already, that $D_i \subset C_i$, therefore $\mathbb{P}\left(B_i\middle|C_i\right) \leq \mathbb{P}\left(B_i\middle|D_i\right)$. We'll show that even this "expanded" probability is small.
For that we use the second inequality from the Fact 1. Analogously $z = s_i$ and t is chosen such that $t\sigma = 1/2 \Rightarrow t = \frac{1}{2\sigma}$. Therefore applying Fact 1:

$$\mathbb{P}(|s_i| \geq 1/2) \leq 2\exp\left(-\frac{t^2}{2}\right) = 2\exp\left(-\frac{1}{8\sigma^2}\right) \tag{21}$$

Recall

$$\sigma^2 = \frac{1}{||X_i||^2}\left(1 + \sum_{j \in \text{supp}\{\beta^*\}\setminus\{i\}} (\beta_j^*)^2\right),$$

and now we need to have a upper bound on σ^2.
We are assuming D_i is happening, therefore similarly to Assignment 6, but with changed estimate on $||X_i||^2$

$$\sigma^2 \leq \frac{10k}{||X_i||^2} \leq \frac{10k}{n/2} \geq \frac{10k}{500k\ln d} \geq \frac{1}{50\ln d} \geq \frac{1}{50 \cdot \ln 1000} \geq 10^{-2} \tag{22}$$

Therefore in (21) we have from

$$\mathbb{P}\left(B_i \middle| C_i\right) \leq \mathbb{P}\left(B_i \middle| D_i\right) \leq 2\exp\left(-\frac{50\ln d}{8}\right) \leq 2\exp\left(-\frac{1}{8\cdot 0.01}\right) = 2\exp\left(-12.5\right) < 2{\cdot}3^{-3^4} < 2{\cdot}10^{-4} < 10^{-3} \tag{23}$$

Returning to (15) and (16) we have

$$\mathbb{P}\left(A_i\right) = \mathbb{P}\left(B_i \middle| C_i\right)\mathbb{P}\left(C_i\right) \leq 2\exp\left(-\frac{50\ln d}{8}\right)\cdot 1 = 2d^{-6.25} \text{ need } \leq \frac{0.01}{d-k} \tag{24}$$

But $\frac{0.01}{d} \leq \frac{0.01}{d-k}$ as $k \geq 1$. Sequence of transformations (of course $d \geq 1000$):

$$\left[2d^{-6.25} \leq \frac{0.01}{d}\right] \leftrightarrow^{*d} \left[2d^{-5.25} \leq 0.01\right] \leftrightarrow^{*100} \left[200d^{-5.25} \leq 1\right] \tag{25}$$

But $d \geq 1000$, therefore $2d^{-5.25} \leq 2\cdot(10^3)^{-5} \leq 2\cdot 10^{15} < 0.01$. From this follows (12), which was needed to prove.

Assignment 8

Use assignments 6 and 7 to prove Theorem 2.

We have $X \in \mathbb{R}^{n\times d}$, $\beta^*, \hat{\beta} \in \mathcal{R}^d$. Need to prove

$$\frac{1}{n}||X\left(\beta^* - \hat{\beta}\right)||^2 = \Omega\left(\frac{k^2}{n}\right) \tag{26}$$

Note that

$$\frac{1}{n}||X\left(\beta^* - \hat{\beta}\right)||^2 = \frac{1}{n}\left(\beta^* - \hat{\beta}\right)^\top X^\top X\left(\beta^* - \hat{\beta}\right) \tag{27}$$

Because when we multiply some matrix $A \in \mathbb{R}^{n\times d}$ and vector $b \in \mathbb{R}^d$, result is a linear combination of columns of matrix with coefficients as corresp. vector elements $\sum_{i=1}^d A_{\cdot i}b_i$, therefore when we multiply by $\left(\beta^* - \hat{\beta}\right)$ we'll have non-zeros only with indices i where $\left(\beta_i^* - \hat{\beta}_i\right)$. Denoting $\delta_i \equiv \beta_i^* - \hat{\beta}_i$ and $S \equiv \mathrm{supp}\{\beta^*\}$. So we can rewrite

$$\frac{1}{n}\delta^\top\left(X^\top X\right)\delta = \frac{1}{n}\sum_{i,j,k}\delta_i^\top\left(X_{j,i}^\top X_{j,k}\right)\delta_k = \frac{1}{n}\sum_{i\in S,j,k\in S}\delta_i^\top\left(X_{j,i}^\top X_{j,k}\right)\delta_k \tag{28}$$

So we basically restricted δ to non-zero elements and X to correspondent columns. We can compile restricted versions δ^* and X^* just by crossing out zeros from δ and corresp. columns from X, and then push remaining elements to be close as in proper vector or matrix.

The size of X^* is $n\times k'$, of δ^* is k', where $k' \leq k$ with probability at least 0.99 (from assignment 7 we have that $\mathrm{s\bar{u}pp}\{\beta^*\} \subseteq \mathrm{s\bar{u}pp}\{\hat{\beta}\}$ with probability at least 0.99, where bar denotes complementary set w.r.t. $[d]$. Therefore at least in $d-k = d-|S|$ indices i we will have $\beta_i^* = \hat{\beta}_i = 0 \Rightarrow \delta_i = 0$). But if we have at least one $i \in S$ such that $\delta_i = 0$, it from Assignment 6 can occur only with probability at most $1 - 0.99 = 0.01$ (because at least one should be zero, and if more, the overall probability of having at least one is only more than 0.99). So, event A s.t. $k' = k$ has probability $P_{k=k'} \geq 0.99$.

If we subtract now matrix $Id_{k'}$ in order to have similar to the Fact 5 construction, the result will be

$$\delta^{*\top}\left(\frac{X^{*\top}X^*}{n} - Id_{k'}\right)\delta^* + \underbrace{\delta^{*\top}\delta^*}_{||\beta^*-\hat{\beta}||^2} \tag{29}$$

$\delta^{*\top}\delta^* = ||\beta^* - \hat{\beta}||^2 \equiv \delta^\top \delta$, because we obtained δ^* from δ just deleting coordinates i where $\delta_i = 0$.

Following the result of the Assignment 6 and hint to this assignment, at least $\frac{\lceil k \rceil}{2}$ coordinates of it comply with the inequality $|\beta_i^* - \hat{\beta}_i| \geq c\sqrt{\frac{k}{n}}$ (event B), so conditioning on $A \cap B$

$$\delta^{*\top}\delta^* = \sum_{i=1}^{k'} \left(|\beta_i^* - \hat{\beta}_i|\right)^2 \geq \frac{\lceil k \rceil}{2} \cdot \Omega\left(\sqrt{\frac{k}{n}}^2\right) \geq \Omega\left(\frac{k^2}{n}\right)$$

as indices that remained in X^* and δ^* are all from S with probability at least 0.99 (event A).
Overall probability now is $P_{A,B} \geq 0.95 \cdot 0.99 > 0.94.$).
Now we would like to show that the first summand in (29) won't spoil this result (e.g. when subtracting the same order).

Using the Fact 5 and $n \geq 1000k \ln d$, $d \geq 1000$ and conditioning on $k' = k$ with prob. at least 0.94, we have for $M \equiv n$, $m \equiv k'$ and $G = X^*$:

$$\mathbb{P}\left(||\frac{X^{*\top}X^*}{n} - Id_{k'}|| \leq 0.9 \middle| k' = k\right) \geq 1-2\exp\left(-\frac{n}{100k'}\right) \geq 1-2\exp\left(-\frac{1000k\ln d}{k}\right) \geq 1-2\exp\left(-1000\ln 1000\right) \tag{30}$$

$$1 - 2\exp\left(-1000\ln 1000\right) \geq 1 - 2\exp(-100) \geq 1 - 10^{-43} \tag{31}$$

Let's name this event C.
So we have an overall probability (A and B doesn't influence this inequality)

$$\mathbb{P}\left(||\frac{X^{*\top}X^*}{n} - Id_{k'}|| \leq 0.9 \middle| A \cap B\right) \cdot P_{A,B} \geq 0.94 \cdot (1 - 10^{-43}) \geq 0.93 \tag{32}$$

Therefore returning to (29) we have with prob. $P_{A,B,C} \geq 0.93$

$$\delta^{*\top}\left(\frac{X^{*\top}X^*}{n} - Id_{k'}\right)\delta^* \leq \left|\left|\frac{X^{*\top}X^*}{n} - Id_{k'}\right|\right| \cdot \delta^{*\top}\delta^* \leq 0.9 \cdot \delta^{*\top}\delta^* \tag{33}$$

and from triangle inequality

$$\delta^{*\top}\left(\frac{X^{*\top}X^*}{n} - Id_{k'}\right)\delta^* + \delta^{*\top}\delta^* \geq \left||\delta^{*\top}\left(\frac{X^{*\top}X^*}{n} - Id_{k'}\right)\delta^*| - |\delta^{*\top}\delta^*|\right| \geq 0.1 \cdot |\delta^{*\top}\delta^*| = \Omega\left(\frac{k^2}{n}\right). \tag{34}$$

with probability at least $P_{A,B,C} \geq 0.93 > 0.9$ what needed.