Optimization for Data Science: Special Assignment 2

Gabriel Hayat

June 16, 2019

Assignment 1. Let $\delta \in [0,1]$ and $k \in [m/2]$. Prove that if β^* is k-block sparse, then with probability at least $1-\delta$

$$\frac{1}{n}||X(\hat{\beta}(\mathbf{y}) - \beta^*)||^2 \leq \mathcal{O}\left(\frac{rk + k\log\frac{m}{k} + \log\frac{1}{\delta}}{n}\right)$$

where $\hat{\beta}(\mathbf{y})$ is the best k-block-selection estimator.

Hint: Follow a similar approach to the Theorem 1.18 in the lecture notes about Regression.

Solution:
Given that the observational model is $\mathbf{y} = X\beta^* + \mathbf{w}$, that $\hat{\beta}(\mathbf{y})$ is the block-lasso estimator and that β^* is part of the optimization domain,

$$||X\hat{\beta}(\mathbf{y}) - \mathbf{y})||^2 \leq ||X\beta^* - \mathbf{y}||^2$$
$$||X(\hat{\beta}(\mathbf{y}) - \beta^*) - \mathbf{w}||^2 \leq ||\mathbf{w}||^2$$

After expanding the term on the left hand side, we get:

$$||X(\hat{\beta}(\mathbf{y}) - \beta^*)||^2 \leq 2\langle \mathbf{w}, X(\hat{\beta}(\mathbf{y}) - \beta^*)\rangle$$

Dividing by $||X(\hat{\beta}(\mathbf{y}) - \beta^*)||$, we get

$$||X(\hat{\beta}(\mathbf{y}) - \beta^*)|| \leq 2\left\langle \mathbf{w}, \frac{X(\hat{\beta}(\mathbf{y}) - \beta^*)}{||X(\hat{\beta}(\mathbf{y}) - \beta^*)||}\right\rangle$$

For each $i \in [m]$, let $X_{S(i)} \in \mathbb{R}^{n\times r}$ be the sub-matrix of X obtained by taking the columns of X with indices in S(i). Then

$$X(\hat{\beta}(\mathbf{y}) - \beta^*) = \sum_{i=1}^{m} X_{S(i)}(\hat{\beta}(\mathbf{y}) - \beta^*)_{S(i)} \tag{1}$$

$$\tag{2}$$

As vector $\hat{\beta}(\mathbf{y}) - \beta^*$ is at most 2k-block sparse, the sum will contain at most 2k non zero vectors.

Let $\Phi_{(T)} \in \mathbb{R}^{n \times 2rk}$ be the orthogonal basis for the column span of all submatrices $X_{S(i)}$ such that $i \in [T]$. Then, similarly to theorem 1.18,

$$\begin{aligned} ||X(\hat{\beta}(\mathbf{y}) - \beta^*)|| &\leq 2 \left\langle \mathbf{w}, \frac{X(\hat{\beta}(\mathbf{y}) - \beta^*)}{||X(\hat{\beta}(\mathbf{y}) - \beta^*)||} \right\rangle \\ &\leq 2 \max_{\substack{T \subset [m] \\ |T|=2k}} \max_{\substack{\mathbf{u} \in \mathbb{R}^{2rk} \\ ||\mathbf{u}||=1}} \langle \mathbf{w}, \Phi_{(T)} u \rangle \\ &= 2 \max_{\substack{T \subset [m] \\ |T|=2k}} ||\Phi_{(T)}^T \mathbf{w}||. \end{aligned}$$

Note that $\Phi_{(T)}^T \mathbf{w} \sim \mathcal{N}(0, Id_{2rk})$ [1].Using the tail bound[2] on its squared Euclidean norm:

$$\mathbb{P}\left\{||\Phi_{(T)}^T \mathbf{w}||^2 \geq t \cdot 2rk\right\} \leq e^{-t \cdot rk/10} \qquad \text{for } t \geq 1$$

Let $t' = 2rk \cdot t$, then

$$\mathbb{P}\left\{||\Phi_{(T)}^T \mathbf{w}||^2 \geq t'\right\} \leq e^{-t'/20} \qquad \text{for } t' \geq 2 \cdot rk \tag{3}$$

By the union bound[3],

$$\mathbb{P}\left\{||X(\hat{\beta}(\mathbf{y}) - \beta^*)||^2 \geq t\right\} \leq \mathbb{P}\left\{\max_{\substack{T \subset [m] \\ |T|=2k}} ||\Phi_{(T)}^T \mathbf{w}||^2 \geq t/2\right\} \tag{4}$$

$$\leq \sum_{\substack{T \subset [m] \\ |T|=2k}} \mathbb{P}\left\{||\Phi_{(T)}^T \mathbf{w}||^2 \geq t/2\right\} \tag{5}$$

$$\leq \binom{m}{2k} \cdot e^{-t/40} \qquad \forall\ t \geq 4 \cdot rk \tag{6}$$

$$\leq \left(\frac{2m}{k}\right)^{2k} \cdot e^{-t/40} \qquad \forall\ t \geq 4 \cdot rk \tag{7}$$

Here, inequality (6) used the tail bound of equation (3) and inequality (7) used the bound $\binom{a}{b} \leq (4a/b)^b$. Suppose:

$$\left(\frac{2m}{k}\right)^{2k} \cdot e^{-t/40} \leq \delta \ \Rightarrow\ t \geq 80k \cdot \log \frac{2m}{k} + 40 \log \frac{1}{\delta}$$

Thus,

$$\mathbb{P}\left\{||X(\hat{\beta}(\mathbf{y}) - \beta^*)||^2 \geq t\right\} \leq \delta \text{ for } t \geq 4 \cdot rk \text{ and } t \geq 80k \cdot \log \frac{2m}{k} + 40 \log \frac{1}{\delta}$$

In particular, take $t = 4 \cdot rk + 80k \cdot \log \frac{2m}{k} + 40 \log \frac{1}{\delta}$ and with probability at least $1 - \delta$:

$$\frac{1}{n}||X(\hat{\beta}(\mathbf{y}) - \beta^*)||^2 \leq \mathcal{O}\left(\frac{rk + k \log \frac{m}{k} + \log \frac{1}{\delta}}{n}\right)$$

□

[1] Regression Lecture notes: Exercise 1.12

[2] Regression Lecture notes: A.21

[3] Regression lecture notes: p.12

Assignment 2. Given $\delta \in [0,1]$, show that with probability at least $1-\delta$,

$$||X^T\mathbf{w}||_{2,\infty}^2 \leq \mathcal{O}\left(r^2 n + rn \cdot \log\frac{m}{\delta}\right)$$

where $\mathbf{w} \sim \mathcal{N}(0, Id_{rn})$
Hint: Use the tail bound on the random variable $\frac{1}{rn}||X_{S(i)}^T\mathbf{w}||^2$.

Solution:
Consider the normally distributed [4] variable $\frac{1}{\sqrt{rn}}X_{S(i)}^T\mathbf{w}$.
The expectation is given by:

$$\mathbb{E}\ \frac{1}{\sqrt{rn}}X_{S(i)}^T\mathbf{w} = \frac{1}{\sqrt{rn}}X_{S(i)}^T\ \mathbb{E}\ \mathbf{w} = 0$$

The variance is given by:

$$\begin{aligned}\mathbb{E}\ \left(\frac{1}{\sqrt{rn}}X_{S(i)}^T\mathbf{w}\right)\left(\frac{1}{\sqrt{rn}}X_{S(i)}^T\mathbf{w}\right)^T &= \mathbb{E}\ \frac{1}{\sqrt{rn}}X_{S(i)}^T\mathbf{w}\mathbf{w}^T X_{S(i)}\frac{1}{\sqrt{rn}} \\ &= \frac{1}{\sqrt{rn}}X_{S(i)}^T\left(\mathbb{E}\ \mathbf{w}\mathbf{w}^T\right)X_{S(i)}\frac{1}{\sqrt{rn}} \\ &= \frac{1}{rn}X_{S(i)}^T X_{S(i)} = \frac{1}{rn}\left(rn \cdot Id_r\right) = Id_r\end{aligned}$$

We conclude that:

$$\frac{1}{\sqrt{rn}}X_{S(i)}^T\mathbf{w} \sim \mathcal{N}(0, Id_r)$$

We use the tail bound[5] on its squared Euclidean norm:

$$\mathbb{P}\left\{\frac{1}{rn}||X_{S(i)}^T\mathbf{w}||^2 \geq t \cdot r\right\} \leq e^{-t\cdot r/10} \qquad \text{for } t \geq 1$$

Let $t' = r^2 n \cdot t$. Then:

$$\mathbb{P}\left\{||X_{S(i)}^T\mathbf{w}||^2 \geq t'\right\} \leq e^{\frac{-t'}{10\cdot rn}} \qquad \text{for } t' \geq r^2 n \tag{8}$$

Therefore,

$$\begin{aligned}\mathbb{P}\left\{||X^T\mathbf{w}||_{2,\infty}^2 \geq t\right\} &= \mathbb{P}\left\{\max_{1\leq i\leq m}||X_{S(i)}^T\mathbf{w}||^2 \geq t\right\} &(9)\\ &= \mathbb{P}\left\{\bigcup_{i=1}^m ||X_{S(i)}^T\mathbf{w}||^2 \geq t\right\} &(10)\\ &\leq \sum_{i=1}^m \mathbb{P}\left\{||X_{S(i)}^T\mathbf{w}||^2 \geq t\right\} &(11)\\ &\leq \sum_{i=1}^m e^{\frac{-t}{10\cdot rn}} = m \cdot e^{\frac{-t}{10\cdot rn}} \qquad \forall\ t \geq r^2 n &(12)\end{aligned}$$

[4] Regression lecture notes: Exercise 1.12
[5] Regression lecture notes: inequality A.21

Inequality (11) is the union bound[6] and inequality (12) use the tail bound of equation (8). Suppose:

$$m \cdot e^{\frac{-t}{10 \cdot rn}} \leq \delta \;\Rightarrow\; t \geq 10 \cdot rn \log \frac{m}{\delta}$$

Thus,

$$\mathbb{P}\left\{||X^T \mathbf{w}||_{2,\infty}^2 \geq t\right\} \leq \delta \text{ for } t \geq r^2 n \text{ and } t \geq 10 \cdot rn \log \frac{m}{\delta}$$

In particular, take $t = r^2 n + 10 \cdot rn \log \frac{m}{\delta}$ and with probability at least $1 - \delta$:

$$||X^T \mathbf{w}||_{2,\infty}^2 \leq \mathcal{O}\left(r^2 n + rn \cdot \log \frac{m}{\delta}\right)$$

□

[6]Regression notes: page 12

Assignment 3. Let $\delta \in [0,1]$. Sow that if $||\beta^*||_{2,1} \leq R$, then with probabilty at least $1-\delta$ the block-lasso estimator $\hat{\beta}(y) = \text{argmin}\,\{||X\beta - y||^2 : ||\beta||_{2,1} \leq R\}$ has mean squared prediction error

$$\frac{1}{rn}||X(\hat{\beta}(y) - \beta^*)||^2 \leq \mathcal{O}\left(\sqrt{\frac{R^2}{n}\left(1 + \frac{log\frac{m}{\delta}}{r}\right)}\right)$$

Hint: Prove that the following variant of Hölder's inequality holds:

$$\mathbf{x}^T\mathbf{y} \leq ||\mathbf{x}||_{2,1} \cdot ||\mathbf{y}||_{2,\infty}$$

Solution: We first demonstrate the hint.
Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{rm}$. By definition of the scalar product:

$$\begin{aligned}
\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{rm} x_i y_i &= \sum_{i=1}^{r} x_i y_i + \sum_{i=r}^{2r} x_i y_i + \cdots + \sum_{i=r(m-1)}^{rm} x_i y_i && (13)\\
&= \mathbf{x}_{S(1)}^T\mathbf{y}_{S(1)} + \mathbf{x}_{S(2)}^T\mathbf{y}_{S(2)} + \ldots \mathbf{x}_{S(m)}^T\mathbf{y}_{S(m)} && (14)\\
&\leq ||\mathbf{x}_{S(1)}|| \cdot ||\mathbf{y}_{S(1)}|| + ||\mathbf{x}_{S(2)}|| \cdot ||\mathbf{y}_{S(2)}|| + \ldots + ||\mathbf{x}_{S(m)}|| \cdot ||\mathbf{y}_{S(m)}|| && (15)\\
&= \sum_{i=1}^{m} ||\mathbf{x}_{S(i)}|| \cdot ||\mathbf{y}_{S(i)}|| && (16)
\end{aligned}$$

Equation (15) used Cauchy-Schwarz inequality. Note that $\mathbf{x}_{S(i)}, \mathbf{y}_{S(i)}$ are defined as in the assignment description.
Define $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$ such that $a_i = ||\mathbf{x}_{S(i)}||$ and $b_i = ||\mathbf{y}_{S(i)}||$. By using the fact that $||\cdot||_1$ and $||\cdot||_\infty$ are dual norms:

$$\begin{aligned}
\sum_{i=1}^{m} ||\mathbf{x}_{S(i)}|| \cdot ||\mathbf{y}_{S(i)}|| = \mathbf{a}^T\mathbf{b} &\leq ||\mathbf{a}||_1 \cdot ||\mathbf{b}||_\infty\\
&= \sum_{i=1}^{m} |a_i| \cdot \max_{1\leq i\leq m} |b_i|\\
&= \sum_{i=1}^{m} ||\mathbf{x}_{S(i)}|| \cdot \max_{1\leq i\leq m} ||\mathbf{y}_{S(i)}|| = ||\mathbf{x}||_{2,1} \cdot ||\mathbf{y}||_{2,\infty}
\end{aligned}$$

which proves the variant of Hölder's inequality. Given that the observational model is $\mathbf{y} = X\beta^* + \mathbf{w}$, that $\hat{\beta}(y)$ is the block-lasso estimator and that β^* is part of the optimization domain, we can write:

$$\begin{aligned}
||X\hat{\beta}(y) - \mathbf{y}||^2 &\leq ||X\beta^* - \mathbf{y}||^2\\
||X(\hat{\beta}(y) - \beta^*) - \mathbf{w}||^2 &\leq ||\mathbf{w}||^2
\end{aligned}$$

After expanding the term on the left hand side, we get:

$$\begin{aligned}
||X(\hat{\beta}(y) - \beta^*)||^2 &\leq 2\langle \mathbf{w}, X(\hat{\beta}(y) - \beta^*)\rangle && (17)\\
&= 2\langle X^T\mathbf{w}, \hat{\beta}(y) - \beta^*\rangle && (18)\\
&= 2 \cdot ||X^T\mathbf{w}||_{2,\infty} \cdot ||\hat{\beta}(\mathbf{y}) - \beta^*||_{2,1} && (19)
\end{aligned}$$

where equation (19) used the variant of Hölder's inequality.
Note that:

$$||\hat{\beta}(\mathbf{y}) - \beta^*||_{2,1} = \sum_{i=1}^{m} ||\hat{\beta}(y)_{S(i)} - \beta^*_{S(i)}|| \tag{20}$$

$$\leq \sum_{i=1}^{m} ||\hat{\beta}(y)_{S(i)}|| + \sum_{i=1}^{m} ||\beta^*_{S(i)}|| \tag{21}$$

$$= ||\hat{\beta}(\mathbf{y})||_{2,1} + ||\beta^*||_{2,1} \leq 2\mathcal{R} \tag{22}$$

Inequality (21) used the triangle inequality and inequality (22) used the bounded norm assumptions from the problem description.
By combining this result and the result obtained in assignment 2, then with probability at least $1 - \delta$:

$$\begin{aligned} ||X\hat{\beta}(y) - y||^2 &\leq 2 \cdot ||X^T\mathbf{w}||_{2,\infty} \cdot ||\hat{\beta}(\mathbf{y}) - \beta^*||_{2,1} \\ &\leq 4 \cdot \mathcal{R} \cdot \mathcal{O}\left(\sqrt{r^2n + rn\ log\frac{m}{\delta}}\right) \\ &= \mathcal{O}\left(\sqrt{\mathcal{R}^2(r^2n + rn\log\frac{m}{\delta})}\right) \end{aligned}$$

Thus the result

$$\frac{1}{rn}||X\hat{\beta}(y) - y||^2 \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{R}^2}{r^2n^2}(r^2n + rn\ log\frac{m}{\delta})}\right) = \mathcal{O}\left(\sqrt{\frac{\mathcal{R}^2}{n}\big(1 + \frac{\log\frac{m}{\delta}}{r}\big)}\right)$$

□

Assignment 4. Show hat there exists a matrix $X \in \mathbb{R}^{n \times rm}$ and a k-block sparse vector $\mathbf{u} \in \mathbb{R}^{rm}$ such that the random vector $\mathbf{z} \in \mathbb{R}^n$ defined as

$$\mathbf{z} = X\mathbf{u} + \mathbf{w}$$

has the same distribution as $\mathbf{y}$, where $\mathbf{w} \sim \mathcal{N}(0, Id_n)$. **Hint**: Note that since the domain of each f_i is $[r]$, regardless of the function, it can be represented as a vector in $\mathbb{R}^r$.

Solution:
Let $\mathbf{x}(i)$ denote the i-th entry of $\mathbf{x}$. Define $X \in \mathbb{R}^{n \times rm}$ and $\mathbf{u} \in \mathbb{R}^{rm}$ as follows:

$$X = \begin{bmatrix} \mathbf{g}(\mathbf{x}_1(1))^T & \mathbf{g}(\mathbf{x}_1(2))^T & \mathbf{g}(\mathbf{x}_1(3))^T & \dots & \mathbf{g}(\mathbf{x}_1(m))^T \\ \mathbf{g}(\mathbf{x}_2(1))^T & \mathbf{g}(\mathbf{x}_2(2)))^T & \mathbf{g}(\mathbf{x}_2(3))^T & \dots & \mathbf{g}(\mathbf{x}_2(m))^T \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{g}(\mathbf{x}_n(1))^T & \mathbf{g}(\mathbf{x}_n(2))^T & \mathbf{g}(\mathbf{x}_n(3))^T & \dots & \mathbf{g}(\mathbf{x}_n(m))^T \end{bmatrix}$$

where

$$g\colon \begin{cases} [r] & \to & \mathbb{R}^r \\ i & \mapsto & \mathbf{x} \quad \text{such that } \mathbf{x}(j) = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j, \end{cases} \end{cases}$$

$$\mathbf{u} = \begin{bmatrix} f_1(1) \\ \vdots \\ f_1(r) \\ f_2(1) \\ \vdots \\ f_2(r) \\ \vdots \\ f_m(1) \\ \vdots \\ f_m(r) \end{bmatrix}$$

f is a k-junta, which means that $m - k$ f_i's are indentically zero, ie:

$$f_i = \begin{bmatrix} f_i(1) \\ \vdots \\ f_i(r) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

Vector $\mathbf{u}$ is thus k-block sparse.

We define $\mathbf{z} \in \mathbb{R}^n$ such that

$$\mathbf{z} = X\mathbf{u} + \mathbf{w}$$

where $\mathbf{w} \sim \mathcal{N}(0, Id_n)$.
By definition of matrix X,

$$\mathbf{z}(i) = \sum_{j=1}^{m} \langle\ \mathbf{g}(\mathbf{x}_i(j)), \mathbf{f}_j\ \rangle + w_i \tag{23}$$

$$= \sum_{j=1}^{m} f_j(\mathbf{x}_i(j)) + w_i = \mathbf{y}(i) \tag{24}$$

where equality (24) was obtained using the definition of function $\mathbf{g}$ and $w_i \sim \mathcal{N}(0, 1)$. By defining the matrix X and vector $\mathbf{u}$ as previously, the random vector $\mathbf{z}$ and $\mathbf{y}$ have the same distribution. □