ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Department of Computer Science
Institute of Theoretical Computer Science
Bernd Gärtner, David Steurer

**Optimization for Data Science** **Special Assignment 2** **FS19**

- The solution is due on **Sunday, June 16th, 2019** by **11:59 pm**. Please send your solution as PDF to `hung.hoang@inf.ethz.ch`. After receiving your file, we will send you a confirmation on the following work day, and at the latest on Monday, June 17th. Make sure you receive this confirmation, otherwise complain timely.
- Please solve the exercises carefully and then write a nice and complete exposition of your solution using a computer, where we strongly recommend to use LaTeX. A tutorial can be found at `http://www.cadmo.ethz.ch/education/thesis/latex`. Handwritten solutions will not be graded!
- For geometric drawings that can easily be integrated into LaTeX documents, we recommend the drawing editor IPE, retrievable at `http://ipe7.sourceforge.net/` in source code and as an executable for Windows.
- Keep in mind the following premises:
  - When writing in English, write short and simple sentences.
  - When writing a proof, write precise statements.

  The conclusion is, of course, that your solution should consist of sentences that are short, simple, and precise!
- This is a theory course, which means: if an exercise does not explicitly say “you do not need to prove your answer” or “justify intuitively”, then a formal proof is **always** required. You can of course refer in your solutions to the lecture notes and to the exercises, if a result you need has already been proved there.
- We would like to stress that the ETH Disciplinary Code applies to this special assignment as it constitutes part of your final grade. The only exception we make to the Code is that we encourage you to verbally discuss the tasks with your colleagues. It is strictly prohibited to share any (hand)written or electronic (partial) solutions with any of your colleagues. We are obligated to inform the Rector of any violations of the Code.
- There will be two special assignments. Both of them will be graded and the average grade will contribute 20% to your final grade. That is, if $S_1$ and $S_2$ are the (unrounded) grades from your respective special assignments and $E$ is the (unrounded) grade from your exam, then your final grade will be $0.1 \cdot S_1 + 0.1 \cdot S_2 + 0.8 \cdot E$, rounded to the nearest quarter (rounding is only applied in this last step). If you do not hand in one of the special assignments, it will be counted with a grade of 1.0.
- As with all exercises, the material of the special assignments is relevant for the exam.

## Block-sparsity

Recall that a vector $\mathsf{x}$ is $k$-*sparse* if it has at most $k$ non-zero entries. In this assignment, we introduce the concept of *block-sparsity.*

For any positive integer $n$, let $[n] = \{1, 2, \dots, n\}$. Let $r$ and $m$ be two positive integers, and let $S(1), \dots, S(m)$ be a partition of $[rm]$ into $m$ pair-wise disjoint sets such that $S(i)$ consists of all integers in the interval $(r(i-1), ri]$. That is $[rm] = \bigcup S(i)$, $|S(i)| = r$ for each $i \in [m]$, and $S(i) \cap S(j) = \emptyset$ for all $i \neq j$. This partition is fixed throughout the assignment.

Given $S \subseteq [rm]$ and $\mathsf{x} \in \mathbb{R}^{rm}$, let $\mathsf{x}_S = (x_j : j \in S)$ be the vector in $\mathbb{R}^{|S|}$ obtained by taking all entries of $\mathsf{x}$ with index in $S$, and concatenating them in the same order to produce a new vector with $|S|$ entries.

For a given vector $\mathsf{x} \in \mathbb{R}^{rm}$, let $B(\mathsf{x}) = |\{j \in [m] : \mathsf{x}_{S(j)} \neq \mathbf{0}_r\}|$, i.e., the number of blocks in which $\mathsf{x}$ has at least one non-zero entry. Given a positive integer $k \leq m$, we say that a vector $\mathsf{x} \in \mathbb{R}^{rm}$ is $k$-*block sparse*, if $B(\mathsf{x}) \leq k$. Notice that $k$-sparsity implies $k$-block sparsity, but $k$-block sparsity guarantees only $kr$-sparsity.

For $n \leq rm$, let $X \in \mathbb{R}^{n \times rm}$ be a design matrix in general position (i.e., every subset of at most $n$ columns is linearly independent), and let $\beta^* \in \mathbb{R}^{rm}$ be a parameter vector. Let $\mathbf{y} = X\beta^* + \mathbf{w}$ be a random vector, with $\mathbf{w} \sim N(0, \mathrm{Id}_n)$. For a realization $y \in \mathbb{R}^n$ of $\mathbf{y}$ and $k \in [m]$, the *best* $k$-*block selection* is the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \|X\beta - y\|^2 \\ \text{subject to} & \beta \in \mathbb{R}^{rm} \text{ being k-block sparse.} \end{array} \tag{1}$$

We refer to the estimator that outputs an optimal solution to this problem as best k-block-selection estimator.

**Assignment 1.** (**20 points**) *Let $\delta \in [0,1]$ and $k \in [m/2]$. Prove that if $\beta^*$ is $k$-block sparse, then with probability at least $1 - \delta$*

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

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

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

## LASSO estimators for k-group sparsity.

Given $\mathsf{x} \in \mathbb{R}^{rm}$ let

$$\|\mathsf{x}\|_{2,1} = \sum_{i=1}^{m} \|\mathsf{x}_{S(i)}\|, \text{ and } \|\mathsf{x}\|_{2,\infty} = \max_{1 \leq i \leq m} \|\mathsf{x}_{S(i)}\|.$$

Let $X \in \mathbb{R}^{rn \times rm}$ be a design matrix. For each $i \in [m]$, let $X_{S(i)}$ be the sub-matrix of $X$ obtained by taking the columns of $X$ with indices in $S(i)$; see Figure 1. Assume throughout this section that for each $i \in [m]$, $X_{S(i)}^\top X_{S(i)} = rnI_r$.

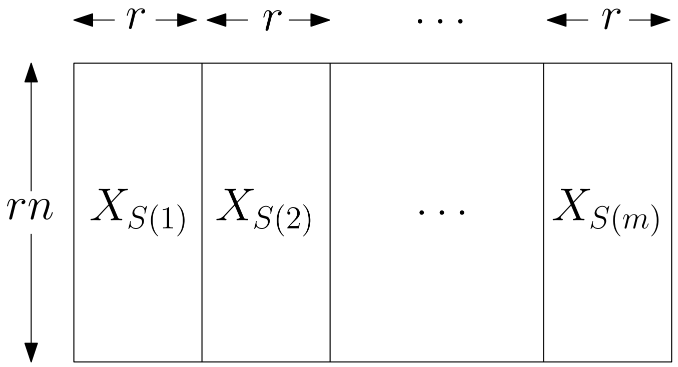

Figure 1: The block structure of X.

**Assignment 2. (15 points)** *Given* $\delta \in [0,1]$*, show that with probability at least* $1-\delta$*,*

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

*where* $\mathbf{w} \sim N(0, \mathrm{Id}_{rn})$*.*

**Hint**: *Use the tail bound on the random variable* $\frac{1}{rn}\|X_{S(i)}^\top \mathbf{w}\|^2$*.*

Let $R > 0$, and let $\beta^* \in \mathbb{R}^{rm}$ be a parameter vector such that $\|\beta^*\|_{2,1} \le R$. Consider the observational model $\mathbf{y} = X\beta^* + \mathbf{w}$, where $\mathbf{w} \sim N(0, \mathrm{Id}_{rn})$. Given a realization $y \in \mathbb{R}^{rn}$ of $\mathbf{y}$, define the *block-lasso estimator*

$$\hat{\beta}(y) = \operatorname{argmin}\{\|X\beta - y\|^2 : \|\beta\|_{2,1} \le R\}.$$

**Assignment 3. (20 points)** *Let* $\delta \in [0,1]$*. Show that if* $\|\beta^*\|_{2,1} \le R$*, then with probability at least* $1-\delta$ *the block-lasso estimator* $\hat{\beta}(y) = \operatorname{argmin}\{\|X\beta - y\|^2 : \|\beta\|_{2,1} \le R\}$ *has mean squared prediction error*

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

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

$$\mathbf{x}^\top \mathbf{y} \le \|\mathbf{x}\|_{2,1} \cdot \|\mathbf{y}\|_{2,\infty}.$$

## Modeling

For this last part of the assignment, we turn to problem modeling.

Suppose we have discrete data points $x_1, \ldots, x_n \in [r]^m$ (sometimes called categorical data) and we observe $n$ real-valued labels for these data points arranged in an $n$-dimensional vector

$y = (y(1), \dots, y(n)) \in \mathbb{R}^n$, where $y(i)$ denotes the i-th entry of $y$. Consider the following model $\mathbf{y}$ for these observations (i.e. $y \sim \mathbf{y}$):

$$\mathbf{y}(i) = f(\mathbf{x}_i) + w_i,$$

where each $w_i \sim N(0, 1)$ is drawn independently. Moreover, $f : [r]^m \to \mathbb{R}$ is of the form $f(\mathbf{x}) = \sum_{j=1}^m f_j(\mathbf{x}(j))$ for $m$ (unknown) functions $f_1, \dots, f_m : [r] \to \mathbb{R}$. We are interested in the case that $f$ is a $k$-junta, which means that at least $m - k$ among the $f_j$'s are identically zero. This implies that $f(\mathbf{x})$ depends on at most $k$ coordinates of $\mathbf{x}$, and so does $\mathbf{y}$. We are interested in capturing this property by using the notion of $k$-block sparsity. To this end, our objective is to define a new random variable, with the same distribution as $\mathbf{y}$, but defined in terms of a $k$-block sparse vector.

**Assignment 4. (15 points)** *Show that 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 $\mathbf{w} \sim N(0, \mathrm{Id}_n)$.*

**Hint**: *Note that since the domain of each $f_i$ is $[r]$, regardless of the function, it can be represented as a vector in $\mathbb{R}^r$.*

By redefining the problem as above, we can find a $k$-block sparse estimator for $u$, and use it to estimate function $f$.