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** **FS20**

- The solution is due on **Sunday, May 31, 2020** by **11:59 pm**. Please submit your solution as PDF on Moodle. The name of the file should follow the format `SA2-{Legi number}`, e.g., `SA2-19-123-456`. After uploading your solution, please make sure that the status is "Submitted for grading". You should receive an automatic email that confirms the submission.

- If you submit your solution within six hours before the deadline and a technical problem prevents you from submitting on Moodle, you can send your solution as PDF to `hung.hoang@inf.ethz.ch`. The cutoff time still needs to be observed. If there is any problem with the submission, 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 and an exam. We will assign percentages to each of them (i.e., the ratio of the points you obtain over the maximum points). The weighted average will then be converted to a grade on the usual scale from 1 to 6. That is, if S_1 and S_2 are the percentages from your respective special assignments and E is the percentage from your exam, then your weighted percentage will be $P = 0.1 \cdot S_1 + 0.1 \cdot S_2 + 0.8 \cdot E$. P will then be converted to your final grade. If you do not hand in one of the special assignments, it will be awarded 0%.
- As with all exercises, the material of the special assignments is relevant for the exam.

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 $w \sim N(0, \mathrm{Id}_n)$ and a k-sparse vector β^* (for some $1 \leqslant k \leqslant d$) such that $\|\beta^*\|^2 \leqslant 10k$ and for any $i \in \operatorname{supp}\{\beta^*\}$, $|\beta^*_i| \geqslant 1$. 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 naïve algorithm, which iteratively applies least square regression to each coordinate. Here, for $i \in [d]$, we denote by X_i the columns of X.

Algorithm 1 (Naïve Sparse Linear Regression).
Input: (y, X)

1. *For all* $i \in [d]$, *if* $X_i \neq \mathbf{0}$, *let* $s_i := \left(X_i{}^\mathsf{T} X_i\right)^{-1} X_i{}^\mathsf{T} y$, *and if* $X_i = \mathbf{0}$, *set* $s_i = 0$.
2. *For all* $i \in [d]$, *if* $|s_i| \geqslant 1/2$ *and* $\|X_i\| \geqslant 10\sqrt{\ln d}$, *let* $\hat{\beta}_i = s_i$. *Otherwise set* $\hat{\beta}_i = 0$.
3. *Return* $\hat{\beta} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_d\right)^\mathsf{T}$.

Our first objective is to prove the following Theorem.

Theorem 1. *Let* $y = X\beta^* + w$ *as in Eq.* (1). *Suppose that* X *satisfies*

$$\forall i \in \operatorname{supp}\{\beta^*\}\ \forall j \in [d] \setminus \{i\}, \quad \langle X_i, X_j\rangle = 0\,. \tag{2}$$

Then, with probability tending to 1 *as* $d \to \infty$, *Algorithm 1 returns vector* $\hat{\beta}$ *such that*

$$\frac{1}{n}\left\|X\left(\beta^* - \hat{\beta}\right)\right\|^2 \leqslant O\left(\frac{k}{n}\cdot \ln d\right). \tag{3}$$

This will encompass Assignment 1, Assignment 2, Assignment 3 and Assignment 4.

Now, given Theorem 1, a natural question is to ask whether the same result can be extended to a larger family of matrices in $\mathbb{R}^{n\times d}$. However, the second theorem we will be proving shows that Algorithm 1 already breaks when X is a standard Gaussian matrix.

Theorem 2. *Let* $d \geqslant 1000$, $n \geqslant 1000k\ln d$, *and let* $y = X\beta^* + w$ *as in Eq.* (1). *Suppose that* $X \sim N(0,1)^{n\times d}$ *and is independent from* w. *Then with probability at least* 0.9, *Algorithm 1 returns vector* $\hat{\beta}$ *such that*

$$\frac{1}{n}\left\|X\left(\beta^* - \hat{\beta}\right)\right\|^2 \geqslant \Omega\left(\frac{k^2}{n}\right). \tag{4}$$

In Assignment 5, Assignment 6, Assignment 7 and Assignment 8 you will prove Theorem 2.

The assignments are listed below. At the end of the document, in Section 3, you will find several statements you can use in your proofs. You can assume such facts to be true and do not need to prove them. However, we may deduct points if you introduce mistakes trying to prove such (or similar) statements.

1 Proof of Theorem 1

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

Assignment 2. **(10 points)** *Consider the settings of Theorem 1. Let* $i \in \operatorname{supp}\{\beta^*\}$. *Show that*

$$\left\|X_i\left(\beta_i^* - \hat{\beta}_i\right)\right\|^2 \leqslant O\left(|\beta_i^*|^2 \ln d\right) \tag{5}$$

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

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* $1 - O\left(d^{-10}\right)$.

Assignment 4. **(10 points)** *Use Assignments 2 and 3 to prove Theorem 1.*

2 Proof of Theorem 2

Assignment 5. **(12 points)** *Consider the settings of Theorem 2. Let* $i \in [d]$. *Show that, given* $\|X_i\|$, *if* $\|X_i\| \neq 0$,

$$s_i - \beta_i^* \sim N\left(0, \sigma^2\right) , \tag{6}$$

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

Assignment 6. **(10 points)** *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*

$$\left|\beta_i^* - \hat{\beta}_i\right| \geqslant c\sqrt{\frac{k}{n}} \tag{7}$$

with probability at least 0.99.

Assignment 7. **(10 points)** *Consider the settings of Theorem 2. Show that with probability at least* 0.99, $\operatorname{supp}\left\{\hat{\beta}\right\} \subseteq \operatorname{supp}\{\beta^*\}$.

Assignment 8. **(13 points)** *Use Assignments 6 and 7 to prove Theorem 2.*

Hint: *You can use the following fact without justification: If* $A_1, \ldots, A_k$ *are random events (possibly dependent) such that for each* $i \in [k]$, $\mathbb{P}(A_i) \geqslant 0.99$, *then with probability at least* 0.95, *at least* $\lceil k/2 \rceil$ *events among* $A_1, \ldots, A_k$ *occur.*

3 Useful Facts

In the solutions you can use the following facts about Gaussian distribution without justifications:

Fact 1. *Let* $z \sim N(0, \sigma^2)$ *for some* $\sigma > 0$. *Then for any* $t \geqslant 0$,

$$1 - 2t \leqslant \mathbb{P}(|z| \geqslant t\sigma) \leqslant 2\exp\left(-t^2/2\right) .$$

Fact 2. *Let* $m \in \mathbb{N}$ *and* $g \sim N(0, \sigma^2 \cdot \mathrm{Id}_m)$ *for some* $\sigma > 0$. *Then* $\frac{1}{\|g\|}g$ *and* $\|g\|$ *are independent.*

Fact 3. *Let* $m \in \mathbb{N}$ *and* $g \sim N(0, \sigma^2 \cdot \mathrm{Id}_m)$ *for some* $\sigma > 0$, *and let* $u \in \mathbb{R}^m$ *be a unit vector independent from* g. *Then* $\langle g, u\rangle \sim N(0, \sigma^2)$.

Fact 4. *Let* $m \in \mathbb{N}$ *and* $g \sim N(0, \mathrm{Id}_m)$. *Then*

$$\mathbb{P}\left(m/2 \leqslant \|g\|^2 \leqslant 2m\right) \geqslant 1 - \exp(-m/100) .$$

Fact 5. *Let* $M, m \in \mathbb{N}$ *and* $G \sim N(0,1)^{M \times m}$. *Then*[1]

$$\mathbb{P}\left(\left\|\frac{1}{M}G^{\mathsf{T}}G - \mathrm{Id}_m\right\| \leqslant 0.9\right) \geqslant 1 - 2\exp\left(-\frac{M}{100m}\right) .$$

[1] The norm in the expression below is the spectral norm.