Exam M2RI 02/2019: Statistics and optimization (3h)

1 Introduction

Most of this problem is based on the lecture notes and on the following article which provides an algorithmic view of sparse recovery.

T. Blumensath & M.E. Davies (2009). Iterative hard thresholding for compressed sensing. Applied and computational harmonic analysis, 27(3), 265-274.

Lecture notes and handwritten notes are allowed. You can answer in english or in french. All sections will be evaluated independantly, you can use results from one section in another section. Some notations are slightly different from what have been seen in class. Comments and interpretation of the results are part of the evaluation.

Notations:

- $\|\cdot\|_0$ denotes the ℓ_0 pseudo norm: the number of nonzero entries of a vector. For $\theta \in \mathbb{R}^d$, $\mathrm{supp}(\theta) \subset \{1,\dots,d\}$ is the set of indices of the nonzero entries of θ. Letting $|\cdot|$ denote the size of a set, we have $\|\theta\|_0 = |\mathrm{supp}(\theta)|$ for any $\theta \in \mathbb{R}^d$.
- For any $S \subset \{1,\dots,d\}$, and $\theta \in \mathbb{R}^d$, we denote by $\theta^S \in \mathbb{R}^d$ the vectors such that $\theta_i = \theta_i^S$ for any $i \in S$ and $\theta_j^S = 0$ for any $j \notin S$.
- For any $S \subset \{1,\dots,d\}$ and $\mathbb{X} \in \mathbb{R}^{n\times d}$, we denote by $\mathbb{X}^S \in \mathbb{R}^{n\times d}$ the matrix which columns indexed by S are the same as in $\mathbb{X}$ and the others are set to 0.
- With these notations, for any $S \subset \{1,\dots,d\}$, denoting by S^c the complement of S in $\{1,\dots,d\}$, we have for any $\theta \in \mathbb{R}^d$

$$\begin{aligned} \mathbb{X}\theta &= \mathbb{X}^S\theta^S + \mathbb{X}^{S^c}\theta^{S^c} \\ \|\theta\|_2^2 &= \|\theta^S\|_2^2 + \|\theta^{S_c}\|_2^2 \\ \mathrm{supp}\left((\mathbb{X}^S)^T x\right) &\subset S, \qquad \forall x \in \mathbb{R}^n. \end{aligned}$$

- In particular if $\mathrm{supp}(\theta) \subset S$, we have $\mathbb{X}\theta = \mathbb{X}^S\theta^S$ and $\|\theta\|_2 = \|\theta^S\|_2$.

1.1 Linear regression with fixed design

Assumption 1.1 (Sparse linear model)**.**

- *Denote by $\mathbb{X} \in \mathbb{R}^{n\times d}$ the design matrix. We have:*

$$Y = \mathbb{X}\theta_* + \epsilon \in \mathbb{R}^n \tag{LM}$$

- *ϵ is a subgaussian random vector with variance proxy $\sigma^2 > 0$.*
- *$\|\theta_*\|_0 \le s$ for some $s \in \mathbb{N}$, $s < d/3$.*

Given the knowledge of $\mathbb{X}$ and $Y \in \mathbb{R}^n$, our goal is to find $\theta \in \mathbb{R}^d$ with small Mean Squared Error:

$$\mathrm{MSE}(\theta) = \frac{1}{n}\|\mathbb{X}\left(\theta - \theta_*\right)\|_2^2.$$

1.2 RIP condition and Iterative Hard Thresholding algorithm

Definition 1.1 (Restricted isometry property (RIP))**.** $\mathbb{X}$ *satisfies the Restricted Isometry Property (RIP) if for all* $\theta \in \mathbb{R}^d$*, with* $\|\theta\|_0 \leq 3s$*, it holds that*

$$\frac{7}{8}\|\theta\|_2^2 \leq \|\mathbb{X}\theta\|_2^2 \leq \|\theta\|_2^2 \tag{1}$$

Algorithm 1.1 (Iterative Hard Thresholding)**.** *Let* $P_s \colon \mathbb{R}^d \mapsto \mathbb{R}^d$ *denotes the projection on the set of* s*-sparse vectors,*

$$P_s(\theta) \in \arg\min_{y \in \mathbb{R}^d} \|y - \theta\|_2^2 \qquad \text{s.t.} \qquad \|y\|_0 \leq s. \tag{2}$$

Given $\mathbb{X}$*,* Y *and* s *as in Assumption 1.1, we consider the following iterative algorithm. Set* $\theta_0 = 0 \in \mathbb{R}^p$ *and iterate for* $k \in \mathbb{N}$*.*

$$\gamma_k = \theta_k - \mathbb{X}^T(\mathbb{X}\theta_k - Y) \tag{3}$$

$$\theta_{k+1} = P_s(\gamma_k) \tag{4}$$

2 Preliminary on restricted isometry property

In this section we assume that $\mathbb{X}$ satisfies the RIP condition as in Assumption 1.1.

2.1 Deterministic results

The RIP condition in Definition 1.1 ensures that the nonzero eigenvalues of $(\mathbb{X}^S)^T\mathbb{X}^S$ are in $[7/8, 1]$ for any $S \subset \{1, \dots, d\}$ such that $|S| \leq 3s$. In this section $S \subset \{1, \dots, d\}$ with $|S| \leq 3s$ is fixed, prove the following

1. For any $x \in \mathbb{R}^n$,
$$\|(\mathbb{X}^S)^T x\|_2 \leq \|x\|_2 \tag{5}$$

2. For any $\theta \in \mathbb{R}^d$,
$$\|(I - (\mathbb{X}^S)^T\mathbb{X}^S)\theta^S\|_2 \leq \frac{1}{8}\|\theta^S\|_2 \tag{6}$$

3. Given $S_2 \subset \{1, \dots, d\}$ such that $|S_2 \cup S| \leq 3s$ and $S_2 \cap S = \emptyset$, for any $\theta \in \mathbb{R}^d$
$$\|(\mathbb{X}^{S_2})^T\mathbb{X}^S\theta^S\|_2 \leq \frac{1}{8}\|\theta^S\|_2. \tag{7}$$
(Hint: Setting $S_3 = S_2 \cup S$, use the fact that $\|(I - (\mathbb{X}^{S_3})^T\mathbb{X}^{S_3})\theta^S\|_2^2 \leq \frac{1}{64}\|\theta^S\|_2^2$).

2.2 RIP and randomness

In this section $\epsilon \in \mathbb{R}^n$ denotes a subgaussian random vector with variance proxy σ^2 as in Assumption 1.1.

4. For any $S \subset \{1, \dots, d\}$ with $|S| \leq 3s$, show that for any $t > 0$,
$$\mathbb{P}\left[\|(\mathbb{X}^S)^T\epsilon\|_2^2 \geq t\right] \leq 6^{|S|}\exp\left(\frac{-t}{8\sigma^2}\right)$$
(Hint: diagonalize the matrix $\mathbb{X}^S(\mathbb{X}^S)^T$).

5. Deduce that for all $t > 0$
$$\mathbb{P}\left[\max_{|S| \leq 3s} \|(\mathbb{X}^S)^T\epsilon\|_2^2 \geq t\right] \leq \binom{d}{3s} 6^{3s}\exp\left(\frac{-t}{8\sigma^2}\right). \tag{8}$$

6. Deduce that for any $\delta > 0$ with probability $1 - \delta$ at least, we have
$$\max_{|S| \leq 3s} \|(\mathbb{X}^S)^T\epsilon\|_2^2 \leq 8\sigma^2\left(3s\log(6d/s) + \log(1/\delta)\right) \tag{9}$$

3 Convergence under RIP condition

This section is devoted to the proof of the following result:

Theorem 3.1. *Under Assumption 1.1, assuming that $\mathbb{X}$ satisfies the RIP condition in Definition 1.1, for any $\delta > 0$, setting $k \geq \max\left\{\log_2\left(\frac{\|\theta_*\|_2^2}{128\sigma^2}\right), 1\right\}$, we have with probability at least $1-\delta$*

$$\mathrm{MSE}(\theta_k) \leq \frac{128\sigma^2\left(1 + 3s\log(6d/s) + \log(1/\delta)\right)}{n}.$$

For any $k \in \mathbb{N}$, we set $S_k = \mathrm{supp}(\theta^*) \cup \mathrm{supp}(\theta_k)$. Note that $|S_k| \leq 2s$ for all $k \in \mathbb{N}$. We assume that $\mathbb{X}$ satisfies the RIP condition in definition 1.1.

7. Using (2), show that for all $k \in \mathbb{N}$

$$\|\theta_* - \theta_{k+1}\|_2 \leq 2\|\gamma_k^{S_{k+1}} - \theta_*^{S_{k+1}}\|_2. \tag{10}$$

8. For all $k \in \mathbb{N}$, using $\gamma_k^{S_{k+1}} = \theta_k^{S_{k+1}} - \left((\mathbb{X}^{S_{k+1}})^T(\mathbb{X}\theta_k - Y)\right)$ deduce that

$$\begin{aligned}\|\theta_* - \theta_{k+1}\|_2 \leq \quad & 2\|(I - (\mathbb{X}^{S_{k+1}})^T\mathbb{X}^{S_{k+1}})(\theta_k^{S_{k+1}} - \theta_*^{S_{k+1}})\|_2 \\ & + 2\|(\mathbb{X}^{S_{k+1}})^T\mathbb{X}^{S_k\setminus S_{k+1}}(\theta_k^{S_k\setminus S_{k+1}} - \theta_*^{S_k\setminus S_{k+1}})\|_2 \\ & + 2\|(\mathbb{X}^{S_{k+1}})^T\epsilon\|_2.\end{aligned} \tag{11}$$

9. Using the results of Sections 2.1 and 2.2, show that for any $\delta > 0$, with probability at least $1-\delta$ over the random draw of ϵ in Assumption 1.1, we have for any $k \in \mathbb{N}$

$$\|\theta_* - \theta_{k+1}\|_2^2 \leq \frac{1}{2}\|\theta_* - \theta_k\|_2^2 + 64\sigma^2\left(3s\log(6d/s) + \log(1/\delta)\right). \tag{12}$$

10. Prove that for any $\delta > 0$, with probability at least $1-\delta$ over the random draw of ϵ in Assumption 1.1, for all $k \in \mathbb{N}$

$$\|\theta_k - \theta_*\|_2^2 \leq \frac{1}{2^k}\|\theta_*\|_2^2 + 128\sigma^2\left(3s\log(6d/s) + \log(1/\delta)\right)$$

11. Prove Theorem 3.1.

12. Comment on the statistical efficiency of the estimator $\hat{\theta} = \theta_k$ with k as in the previous question. Make connections with the estimators seen in class.

4 RIP and sparse integral solution to linear systems

We consider the following decision problem: given $\mathbb{X}$ and Y fixed, find $\theta \in \mathbb{Z}^d$ with $\|\theta\|_0 \leq s$ such that $\mathbb{X}\theta = Y$. We are interested in the following theorem

Theorem 4.1. *There exists an algorithm such that, if $\mathbb{X}$ satisfies the RIP condition and $Y = \mathbb{X}\theta_*$ with $\theta_* \in \mathbb{Z}^d$, $\|\theta_*\|_0 \leq s$, then the algorithm computes θ_* based on the input $\mathbb{X}$ and Y in polynomial time.*

13. We have seen in class a hardness result for these types of decision problems. Recall this hardness result and comment on what it implies regarding the possibility to solve the decision problem efficiently.

14. Let $s \in \mathbb{N}$ is as in Assumption 1.1 and $\theta \in \mathbb{R}^d$. Projecting θ on the set of s-sparse vectors amounts to solve the optimization problem (2). Decribe a polynomial time algorithm to compute $P_s(\theta)$.

15. Assuming that $\mathbb{X}$ satisfies the RIP condition as in Assumption 1.1, use (11) and the analysis in Section 2.1 to show that if a solution θ_* to the decision problem exists, then after k iteration of the hard thresholding algorithm.

$$\|\theta_k - \theta_*\|_2 \leq \frac{1}{2^k}\|\theta_*\|_2.$$

16. Prove Theorem 4.1.

17. Explain why this is not in contradiction with the first question of this section. Make connections with estimators seen in class. Comment on the interplay between statistical and algorithmic efficiency.

5 A descent algorithm

In the following, we set

$$\begin{aligned} f\colon \mathbb{R}^d &\mapsto \mathbb{R} \\ \theta &\mapsto \frac{1}{2}\|\mathbb{X}\theta - Y\|_2^2 \end{aligned}$$

where $\mathbb{X}$ and Y are given in (LM). The goal of this section is to prove that f is decreasing along the sequence generated by the iterative hard thresholding algorithm.

18. Prove the following identity, for any $\alpha, \beta \in \mathbb{R}^d$:

$$\frac{1}{2}\alpha\mathbb{X}^T\mathbb{X}\alpha + \alpha^T\mathbb{X}^T\mathbb{X}(\beta - \alpha) - \frac{1}{2}\beta^T\mathbb{X}^T\mathbb{X}\beta = -\frac{1}{2}\left(\beta - \alpha\right)^T\mathbb{X}^T\mathbb{X}\left(\beta - \alpha\right)$$

19. Deduce that if $\mathbb{X}$ satisfies the RIP property in Definition 1.1, we have for all $k \in \mathbb{N}$:

$$\begin{aligned} \frac{1}{2}\theta_{k+1}^T\mathbb{X}^T\mathbb{X}\theta_{k+1} &\leq \frac{1}{2}\theta_k^T\mathbb{X}^T\mathbb{X}\theta_k + \theta_k^T\mathbb{X}^T\mathbb{X}(\theta_{k+1} - \theta_k) + \frac{1}{2}\left\|\theta_{k+1} - \theta_k\right\|_2^2 \\ f(\theta_{k+1}) &\leq f(\theta_k) + \left(\mathbb{X}^T\left(\mathbb{X}\theta_k - Y\right)\right)^T\left(\theta_{k+1} - \theta_k\right) + \frac{1}{2}\left\|\theta_{k+1} - \theta_k\right\|_2^2 \end{aligned}$$

20. Combine the preceeding question with the identity given in (2) to show that f is decreasing along the sequence $(\theta_k)_{k\in\mathbb{N}}$. Which result of the course does this remind you? What is the difference?