High Dimensional Learning
Dimensionality reduction

Master 2 SIF

Aline Roumy

January 2022

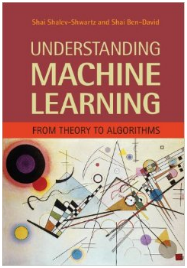
About me

Aline Roumy
Researcher at Inria, Rennes
Expertise: compression for video streaming
image/signal processing, information theory, machine learning

Web: `http://people.rennes.inria.fr/Aline.Roumy/`
email: `aline.roumy@inria.fr`

Course material

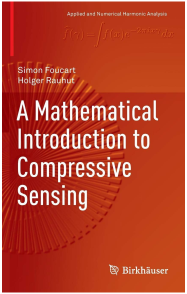
Shai Shalev-Shwartz and Shai Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.



Website and online version at **(web)**

Course material

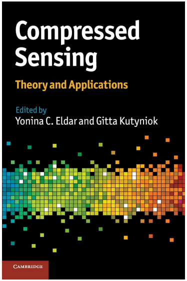
S. Foucart, H. Rauhut, A mathematical introduction to compressive sensing, Birkhaüser, 2013.



Early and short version:
S. Foucart, Notes on compressed sensing, 2009. (pdf)

Course material

Compressed Sensing: Theory and Applications, Edited by Y.C. Eldar and G. Kutyniok, Cambridge University Press, 2012.



- Chapter 1:
  M.A. Davenport, M.F. Duarte, Y.C. Eldar, G. Kutyniok Introduction to compressed sensing. **(pdf)**
- Short version:
  G. Kutyniok, Theory and Applications of Compressed Sensing, GAMM Mitteilungen 36 (2013), 79-101.

Lecture 3 - LINEAR dimensionality reduction and NON-LINEAR reconstruction = Compressive sensing

1. 3.1. Reconstruction guarantee: Restricted Isometry Property
2. 3.2 Iterative Hard Thresholding satisfies RIP: IHT $\Rightarrow$ RIP
3. 3.3. Which matrices satisfy the RIP?
4. 3.4. Summary on Compressive sensing

Reconstruction guarantee: Restricted Isometry Property (RIP)

The problem: invert $y = Mx$

M square 

$\exists$ a reconstruction map:

$$\mathbb{R}^m \to \mathbb{R}^d$$

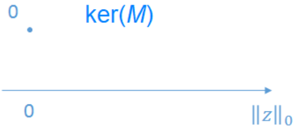
$$y \mapsto x = M^{-1}y$$

$$\Updownarrow$$

condition on the matrix

$$\text{rank}(M) = m = d$$

$$\ker(M) = \{z : Mz = 0\} = \{0\}$$



The problem: invert $y = Mx$

M square 

$\exists$ a reconstruction map:

$$\mathbb{R}^m \to \mathbb{R}^d$$

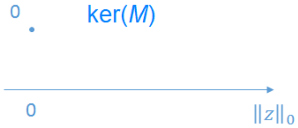
$$y \mapsto x = M^{-1} y$$

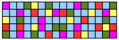
$$\Updownarrow$$

condition on the matrix

$$\text{rank}(M) = m = d$$

$$\ker(M) = \{z : Mz = 0\} = \{0\}$$



M fat 

$\exists$ a reconstruction map:

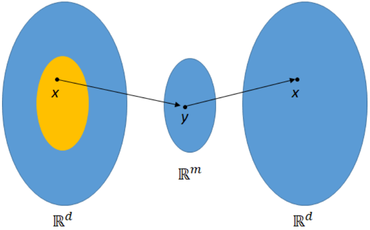
$$\mathbb{R}^m \to \mathbb{R}^d$$

$$y \mapsto x = ??$$

$$\Updownarrow$$

condition on the matrix ???

NEW Reduce the domain of definition of M: s-sparse



The restricted isometry property (RIP): definition

Definition (RIP)

Let $\epsilon > 0$, $s, m, d \in \mathbb{N}$. A matrix $M \in \mathbb{R}^{m,d}$ with $m \leq d$ is (ϵ, s)-RIP if

$$\forall x \in \Sigma_s, \ (1-\epsilon)||x||_2^2 \leq ||Mx||_2^2 \leq (1+\epsilon)||x||_2^2 \quad (1)$$

Interpretation of (ϵ, s)-RIP:

- M preserves the Euclidean norm of s-sparse vectors

The restricted isometry property (RIP): definition

Definition (RIP)

Let $\epsilon > 0$, $s, m, d \in \mathbb{N}$. A matrix $M \in \mathbb{R}^{m,d}$ with $m \leq d$ is (ϵ, s)-RIP if

$$\forall x \in \Sigma_s, \; (1-\epsilon)||x||_2^2 \leq ||Mx||_2^2 \leq (1+\epsilon)||x||_2^2 \tag{1}$$

Interpretation of (ϵ, s)-RIP:

- M preserves the Euclidean norm of s-sparse vectors
- $M \in \mathbb{R}^{m,d}$ with $m \ll d$ is (ϵ, s)-RIP if

$$\forall x \in \Sigma_s \setminus \{0\}, \; \left| \frac{||Mx||_2^2 - ||x||_2^2}{||x||_2^2} \right| \leq \epsilon \tag{2}$$

The restricted isometry property (RIP): definition

Definition (RIP)

Let $\epsilon > 0$, $s, m, d \in \mathbb{N}$. A matrix $M \in \mathbb{R}^{m,d}$ with $m \leq d$ is (ϵ, s)-RIP if

$$\forall x \in \Sigma_s, \ (1-\epsilon)||x||_2^2 \leq ||Mx||_2^2 \leq (1+\epsilon)||x||_2^2 \tag{1}$$

Interpretation of (ϵ, s)-RIP:

- M preserves the Euclidean norm of s-sparse vectors
- $M \in \mathbb{R}^{m,d}$ with $m \ll d$ is (ϵ, s)-RIP if

$$\forall x \in \Sigma_s \setminus \{0\}, \ \left| \frac{||Mx||_2^2 - ||x||_2^2}{||x||_2^2} \right| \leq \epsilon \tag{2}$$

Quiz 1, 2

RIP: l_0 reconstruction

Proposition (RIP and l_0 reconstruction)

Let $M \in \mathbb{R}^{m,d}$ with $m \ll d$. Let $0 < \epsilon < 1$. If M is $(\epsilon, 2s)$-RIP, then

$$\forall x \in \Sigma_s, \quad \hat{x} = x, \text{ with } \hat{x} \in \arg \min_{z : Mz = y} ||z||_0.$$

Proof. *Blackboard + Th 2.13 of Foucart-Rauhut.*

Interpretation:
"a $(\epsilon, 2s)$-RIP matrix is a good sensing matrix for l_0 reconstruction.' We "pay" $2s$ instead of s, because the support is unknown.

THE Question: *is a $(\epsilon, 2s)$-RIP matrix a good sensing matrix for practical reconstruction algorithms?*

RIP: operator norm

Lemma (RIP and operator norm)

Let $M \in \mathbb{R}^{m,d}$ with $m \ll d$. If M is (ϵ, s)-RIP, then

$$\forall S \subset [\![1, d]\!], |S| \leq s, \quad ||M_S^T M_S - I_S||_{op} \leq \epsilon. \tag{3}$$

Recall (note in the definition below $||.||_2$ not $||.||_2^2$)

$$||M_S^T M_S - I||_{op} = \sup_{x_S \neq 0} \frac{||(M_S^T M_S - I)x_S||_2}{||x_S||_2}.$$

Proof. **Quiz 3**

Interpretation: **Quiz 4**

$\forall S, M_S^T M_S \approx I_S$ when applied to any x_S (vector of size S)

A practical algorithm Iterative Hard Thresholding satisfies RIP

Iterative Hard Thresholding (IHT) $\Rightarrow$ *RIP*

A practical algorithm

Definition (Iterative Hard Thresholding (IHT))

$$x^0 = 0$$
$$x^{l+1} = H_s\left(x^l + M^T(y - Mx^l)\right)$$
$$\text{output: } \hat{x}_{IHT} = \lim_{l\to+\infty} x^l$$

H_s: Hard Thresholding
keeps the s coefficients with largest absolute value.

Justification: $x^{l+1} = H_s(x^l + \underbrace{\text{error}(y, M, x^l)}_{\approx x - x^l})$

RIP is good for IHT

Theorem (Optimality of IHT for RIP matrices)

Let $M \in \mathbb{R}^{m,d}$ with $m \ll d$. Let $\epsilon > 0$.
If M is $(\epsilon, \mathbf{3}s)$-RIP, then

$$||x^{l+1} - x|| \leq 2\epsilon ||x^l - x|| \tag{4}$$

Interpretation: **Quiz 5**

RIP is good for IHT

Theorem (Optimality of IHT for RIP matrices)

Let $M \in \mathbb{R}^{m,d}$ *with* $m \ll d$. *Let* $\epsilon > 0$.
If M is $(\epsilon, 3s)$*-RIP, then*

$$||x^{l+1} - x|| \leq 2\epsilon ||x^l - x|| \tag{4}$$

In particular, if $\epsilon < \frac{1}{2}$, $\quad x^l \xrightarrow[l \to +\infty]{} x.$

Interpretation: **Quiz 5**

Proof. **Quiz 6**

Summary: if M is $(\epsilon, 3s)$-RIP, with $\epsilon < 1/2$, then $\hat{x}_{IHT} = x$

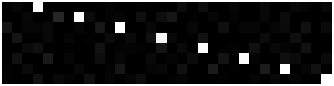
Similarly: if M is $(\epsilon, 2s)$-RIP, with $\epsilon < 1/3$, then $\hat{x}_{BP} = x$ [FR, Th 6.9]

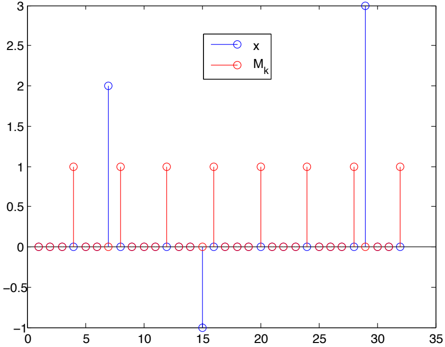
if M is $(\epsilon, 13s)$-RIP, with $\epsilon < 1/6$, then $\hat{x}_{OMP} = x$ [FR, Th 6.25]

THE question: how to construct a matrix M that is $(1/2, 3s)$-RIP?

Which matrices satisfy the RIP?

Sensing matrices that are not good

$M =$ 



Vector $y = Mx$ is all zero!
$\rightarrow$ If x sparse, M must be **non-sparse**

Concentration inequality

Theorem (Concentration of Gaussian Matrices [UML Lemma B.12]**)**

Let $x \in \mathbb{R}^d$. *Let* $M \in \mathbb{R}^{m,d}$ *s.t.* $M_{i,j} \sim \mathcal{N}(0, 1/m)$ *i.i.d.*

$$\forall\, 0 \le t \le 3, \quad \mathbb{P}_M\left(\underbrace{\left|\frac{||Mx||_2^2}{||x||_2^2} - 1\right| > t}_{(*)}\right) \le 2e^{-\frac{mt^2}{6}} \tag{5}$$

Interpretation:

- $\text{Neg}(*) \Leftrightarrow (1-t)||x||_2^2 \le ||Mx||_2^2 \le (1+t)||x||_2^2 \Leftrightarrow M$ is good for this x
- **Quiz 7**

Concentration inequality

Theorem (Concentration of Gaussian Matrices [UML Lemma B.12]**)**

Let $x \in \mathbb{R}^d$. *Let* $M \in \mathbb{R}^{m,d}$ *s.t.* $M_{i,j} \sim \mathcal{N}(0, 1/m)$ *i.i.d.*

$$\forall\ 0 \le t \le 3,\quad \mathbb{P}_M\left(\underbrace{\left|\frac{||Mx||_2^2}{||x||_2^2} - 1\right| > t}_{(*)}\right) \le 2e^{-\frac{mt^2}{6}} \tag{5}$$

Interpretation:

- $\text{Neg}(*) \Leftrightarrow (1-t)||x||_2^2 \le ||Mx||_2^2 \le (1+t)||x||_2^2 \Leftrightarrow M$ is good for this x
- **Quiz 7**

$$(5)\ \Leftrightarrow \underbrace{\exists \alpha, \delta \text{ s.t. } \mathbb{P}_M\left(\left|||Mx||_2^2 - \mathbb{E}[||Mx||_2^2]\right| > \alpha\right) \le \delta}_{\text{concentration (around the mean) inequality}}$$

Other concentration inequalities

Markov's inequality (due to Chebyshev (Markov's teacher)):
Given a non-negative random variable X with finite mean

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}, \quad \forall t > 0. \qquad \text{Decay in } \mathcal{O}(\frac{1}{t}) \tag{6}$$

Other concentration inequalities

Markov's inequality (due to Chebyshev (Markov's teacher)):
Given a non-negative random variable X with finite mean

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}, \quad \forall t > 0. \qquad \text{Decay in } \mathcal{O}(\frac{1}{t}) \tag{6}$$

Chebyshev's inequality: Given a random variable X with mean μ and finite variance (denoted $\text{var}(X) < \infty$)

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\text{var}(X)}{t^2}, \quad \forall t > 0. \qquad \text{Decay in } \mathcal{O}(\frac{1}{t^2}) \tag{7}$$

Other concentration inequalities

Chernoff bound: (due to Herman Rubin)
Given a random variable X with mean μ and finite variance

$$\mathbb{P}(X - \mu \geq t) \leq \frac{\mathbb{E}[e^{\lambda|X-\mu|}]}{e^{\lambda t}}, \quad \forall t, \lambda > 0. \text{ Decay in } \mathcal{O}(e^{-\lambda t}) \tag{8}$$

Other concentration inequalities

Chernoff bound: (due to Herman Rubin)
Given a random variable X with mean μ and finite variance

$$\mathbb{P}(X - \mu \geq t) \leq \frac{\mathbb{E}[e^{\lambda|X-\mu|}]}{e^{\lambda t}}, \quad \forall t, \lambda > 0. \text{ Decay in } \mathcal{O}(e^{-\lambda t}) \tag{8}$$

Cramer-Chernoff method:

step 1 Apply Chernoff bound

step 2 Bound optimization

$$\inf_{\lambda>0} \frac{\mathbb{E}[e^{\lambda|X-\mu|}]}{e^{\lambda t}}$$

step 3 Repeat with $X' := -X$.

Difference between RIP and concentration

Concentration inequality for Gaussian matrices (5) means
Given x

$$\mathbb{P}_M \left(\left| ||Mx||_2^2 - ||x||_2^2 \right| > t||x||_2^2 \right) \leq 2e^{-\frac{mt^2}{6}}$$

RIP means
For all x s-sparse

$$(1-t)||x||_2^2 \leq ||Mx||_2^2 \leq (1+t)||x||_2^2$$

Quiz 8

Condition for “RIP” over FINITE set

Lemma (Johnson-Lindenstrauss)

Let $M \in \mathbb{R}^{m,d}$ *s.t.* $M_{i,j} \sim \mathcal{N}(0, 1/m)$. *Let* $0 \leq t \leq 3, \ \delta > 0$.
Let $\mathcal{Q}$ *be a* **finite set** *of vectors* $\subset \mathbb{R}^n$.

If $m \geq \frac{6}{t^2} \log \frac{2|\mathcal{Q}|}{\delta}$, *then*

$$\mathbb{P}_M \left(\sup_{x \in \mathcal{Q}} \left| \frac{||Mx||_2^2}{||x||_2^2} - 1 \right| \leq t \right) \geq 1 - \delta \tag{9}$$

Interpretation: with probability at least $1 - \delta$, the norm of the vectors is preserved (precision t).

Proof: **Quiz 9**

Condition for RIP and success of IHT

Theorem (RIP and success of IHT [FR, Th. 6.15 and Chap. 12.5]**)**

Let $M \in \mathbb{R}^{m,d}$ s.t. $M_{i,j} \sim \mathcal{N}(0, 1/m)$. Let $\epsilon > 0, \delta > 0$.
If $m \geq \frac{4}{\epsilon^2}\left(2s \ln \frac{en}{s} + 7s + 2 \ln \frac{2}{\delta}\right)$, then

$$\mathbb{P}_M \left(\sup_{x \in \Sigma_s} \left| \frac{||Mx||_2^2}{||x||_2^2} - 1 \right| > \epsilon \right) \leq \delta \tag{10}$$

In particular: $\exists c_1, c_2, c_3 > 0$ s.t. if $m \geq c_1 s \ln \frac{n}{s} + c_2 s + c_3$,
then with probability at least $1 - \delta$

$$\hat{x}_{IHT} = x$$

Proof: **Quiz 10**

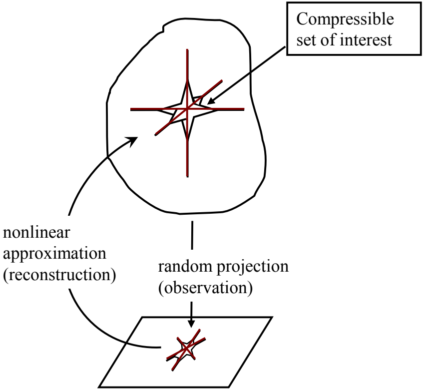
3.4. Summary on Compressive sensing

Compressive sensing overview

Observe $x \in \mathbb{R}^d$ via m measurements, with $m \ll d$
More precisely, $y = Mx$ where $y \in \mathbb{R}^m$

Assumptions:

- signal approximately s-sparse
- use $m \geq c \; s \log \frac{n}{s}$, c=constant, random linear measurements
- reconstruct by a non linear mapping



Mike Davies.