

# Sketching : A framework for compressive learning

Leo Davy   Martin Gjorgjevski   Aleksandr Pak

ENS Lyon  
M2 Advanced Mathematics

March 2022

# Motivation and principles of compressive learning

## What is the setup?

We are confronted with a dataset which comes in form of  $n$   $d$ -dimensional vectors  $\{x_i\}_{i=1}^n$ . We would like to perform some kind of learning on it but we are scared of the complexity when  $n$  is huge.

## What is compressive learning?

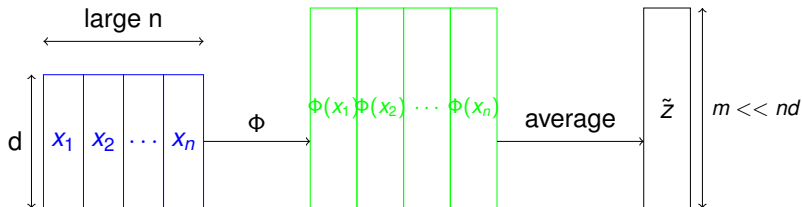
The principle of compressive learning consists of compressing (sketching) the dataset before applying any learning techniques. The sketch consists of a single vector  $\tilde{z}$  which is constructed by transforming each vector of the dataset and averaging the results :

$$\tilde{z} = \frac{1}{n} \sum_{i=1}^n \Phi(x_i)$$

# Motivation and principles of compressive learning

$x_1$	$x_2$	$\dots$	$x_n$
$x_{11}$	$x_{12}$	$\dots$	$x_{1n}$
$x_{21}$	$x_{22}$	$\dots$	$x_{2n}$
$\vdots$	$\vdots$	$\dots$	$\vdots$
$x_{d1}$	$x_{d2}$	$\dots$	$x_{dn}$

Table: Initial dataset



# Principal Component Analysis (PCA)

## PCA

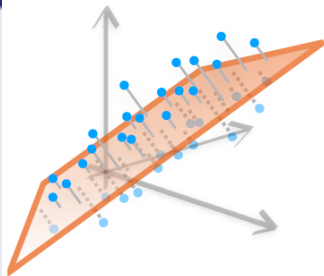
The goal is to find the linear subspace  $P_k$  that best fits the  $d$ -dimensional data  $\{x_i\}_{i=1}^N$  in the LS sense, i.e., find an orthogonal family of  $k$  vectors  $\{u_l\}_{l=1}^k$  that maximizes

$$\sum_{l=1}^k \sum_{i=1}^N |u_l^T x_i|^2.$$

A solution is the  $k$ -principal eigenvectors of the empirical autocorrelation matrix

$$\hat{R} = \frac{1}{N} \sum_{i=1}^N x_i x_i^T =: \frac{1}{N} \sum_{i=1}^N \Phi(x_i).$$

$\hat{R}$  is a *sketch* of our data (of dim  $d^2$ ).



The sketch

$$\hat{R} = \frac{1}{N} \sum_{i=1}^N x_i x_i^T =: \frac{1}{N} \sum_{i=1}^N \Phi(x_i) \in \mathbb{R}^{d^2}$$

is a very compressed version of the data  $\{x_i\}_{i=1}^N$ , *but*, it still contains the geometry of the data.

### CS inspired idea

Take  $m$  random measurements<sup>a</sup> of each sample and use the sketch defined by  $\Phi(x) = \mathcal{M}(xx^T)$ . Provided  $m > kd$ , the principal eigenvectors can be recovered.

---

<sup>a</sup> $\mathcal{M} : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}^m$  satisfying RIP on matrices of rank at most  $2k$ .

# k-means centroids

## The problem

The goal is to recover  $k$  centroids  $\{c_l\}_{l=1}^k$  from some data  $\{x_i\}_{i=1}^N$  that minimize

$$\sum_{i=1}^N \min_l \|x_i - c_l\|^2.$$

For  $N \gg 1$  traditional algorithms are not very efficient because they take the whole dataset at once...

*But*,  $N \gg 1$  allows to use the laws of large numbers and concentration. It is reasonable to consider that the data will accumulate on small portions of the space.

## The binning map

Assume the centroids are spaced by at least  $\varepsilon$  and have a norm smaller than  $r$ , then cover  $[-r, r]^d$  by,  $B = (\frac{2r}{\varepsilon})^d$ ,  $d$ -dimensional cubes (*bins*). For each bin, count the average number of points that belong to it. This defines the binning map  $\hat{p} \in \mathbb{R}_+^B$ .

This gives us a sketch of the data, but in a large dimensional space.

*But*, if the model that generated the data is "structured", i.e., the data concentrates in a few centroids, then the problem is a *sparse* problem.

## CS inspired idea

Use a Gaussian random matrix in  $\mathbb{R}^{m \times N}$  and define the sketch as  $\tilde{z} = A\hat{p}$  and solve

$$\tilde{p} = \operatorname{argmin}_{p \in \Sigma_k^+} \|\tilde{z} - Ap\|^2$$

# Sketching estimates the underlying data-distribution

We can do  $k$ -means clustering with  $m$  measures using:

- $m \geq k \log B$  for Gaussian sampling
- $m \geq k \log(k)^3 \log B$  for DFT of  $\hat{p}$

What if we consider the *continuous* Fourier Transform (FT) ?

We only know the empirical distribution  $\bar{p}_{\mathcal{X}} = \frac{1}{N} \sum_{i=1}^N \delta(x_i - x)$ ,  
so

$$FT(\bar{p})(\omega) = \int_{\mathbb{R}^d} \bar{p}_{\mathcal{X}}(x) e^{-i2\pi \langle w, x \rangle} dx = \frac{1}{N} \sum_{i=1}^N e^{-i2\pi \langle w, x_i \rangle} =: \bar{\Psi}_{\bar{p}_{\mathcal{X}}}(\omega)$$

$\longrightarrow^{\mathbb{E}} \bar{\Psi}_{p^*}(\omega)$  The "true" characteristic function at  $\omega$ .

## CS inspired idea

If the true distribution  $p^*$  is "simple", then interpolating the characteristic function, using simple models, on "few" of its samples should give the true distribution.



# Parallel with signal processing

- Recall that in signal processing our goal is to reconstruct a vector  $x \in \mathbb{R}^d$  from  $y = Ax + \epsilon$
- At first glance compressive learning setup is rather different as we deal with a large collection of vectors, rather than just one
- The analogy becomes clearer if we assume (as it is often the case in ML) that our vectors  $\{x_i\}_{i=1}^n$  are modeled as i.i.d. random vectors having a probability measure  $\mathbb{P}$
- In this case we get

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \stackrel{\text{a.s.}}{=} \mathbb{E}_{\mathbb{P}}[\Phi(X)] = \mathcal{A}(\mathbb{P}),$$

where  $\mathcal{A}$  is a linear operator matching a probability measure to the expectation over this measure of the feature map  $\Phi$

# Parallel with signal processing

- In this manner we can write

$$\tilde{z} = \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \approx \mathcal{A}(\mathbb{P}) = \mathcal{A}(\mathbb{P}) + \epsilon$$

- Thus instead of considering a linear projection of a vector measured with noise ( $x \rightarrow Ax + \epsilon$ ), in compressive learning we consider a linear projection of the underlying probability measured with noise:  $\mathbb{P} \rightarrow \mathcal{A}(\mathbb{P}) + \epsilon$
- In signal processing the linear measurement matrix  $A$  can be chosen at random to ensure good reconstruction properties with high probability. By analogy, in compressive sensing  $\Phi$  is also often randomised

# Task driven distances

## Task driven distance

Given a loss function  $L$  and probability distributions  $p_X$  and  $p'_X$ , we consider the distance

$$\rho(p_X, p'_X) = \sup_{\theta} |R^*(\theta|p_X) - R^*(\theta|p'_X)|$$

where  $R^*(\theta|p_X) = E_{X \sim p_X}(L(\theta)|X)$  is the expected risk under  $p_X$ .

- The loss function  $L$  is task specific
- Excess risk bounds  
 $0 \leq R^*(\theta^*'|p_X) - R^*(\theta^*|p_X) \leq 2\rho(p_X, p'_X)$  where  
 $\theta^* = \operatorname{argmin}_{\theta} R^*(\theta|p_X)$  and  $\theta^*' = \operatorname{argmin}_{\theta} R^*(\theta|p'_X)$
- When  $\hat{p}_X$  is the empirical distribution on the data, and  $p_X$  is the true distribution, under certain conditions  
 $\rho(\hat{p}_X, p_X) = O(\frac{1}{\sqrt{n}})$ .

In the compressive learning framework we are interested in an upper bound of the excess risk which is controlled by the task driven distance.

### The Lower Restricted Isometry Property (LRIP)

The operator  $\mathcal{A}$  is said to have the LRIP with constant  $C_0$  and with respect to a parametric subfamily  $\Sigma_\theta = \{p_\theta | \theta \in \Theta\}$  if

$$\rho(p_\theta, p_{\theta'}) \leq C_0 \|\mathcal{A}(p_\theta) - \mathcal{A}(p_{\theta'})\|$$

for all  $p_\theta, p_{\theta'} \in \Sigma_\theta$

Excess risk bound under LRIP: for all  $\theta' \in \Sigma_\theta$

$$\begin{aligned} R^*(\tilde{\theta} | p_X) - R^*(\theta' | p_X) &\leq 4C_0 \|\mathcal{A}(p_X) - \tilde{z}\| \\ &\quad + 4C_0 \|\mathcal{A}(p_{\theta'}) - \mathcal{A}(p_X)\| + 2\rho(p_{\theta'}, p_X) \end{aligned}$$

with  $\tilde{\theta} = \operatorname{argmin}_\theta C(\theta | \tilde{z})$

- Choosing  $\theta' = \theta^*$ , this result is interpretable in terms of modeling and sampling error

# Expected and mean kernel, MMD

- Two sources of randomness: the data and the random features used for the sketch
- For the random feature map we have
$$\langle \frac{1}{m}\Phi(x), \frac{1}{m}\Phi(x') \rangle = \frac{1}{m} \sum_{j=1}^m e^{-j2\pi w_j(x-x')}$$
- Averaging over the random features gives the *expected kernel*  $k(x, x') = E_w(\exp(-j2\pi \langle w, x - x' \rangle))$
- We define the *mean kernel* as  $k(p, q) = E_{x \sim p, x' \sim q} k(x, x')$
- A quantity of interest is the *maximum mean discrepancy*
$$MMD(p, q) = \sqrt{k(p, p) - 2k(p, q) + k(q, q)}$$
- It can be shown using concentration of measure that
$$\frac{1}{\sqrt{2}} MMD(p_\theta, p_{\theta'}) \leq \frac{1}{\sqrt{m}} \|\mathcal{A}(p_\theta) - \mathcal{A}(p_{\theta'})\|$$
 when  $\Sigma_\theta$  is a finite set
- When  $\Sigma_\theta$  is infinite additional assumptions are required to ensure that the LRIP property holds

# Compressed clustering, fast sketching

- In practice computing the sketch using random Fourier samples might be problematic (accurate implementations of the complex valued function  $x \rightarrow \exp(-j2\pi x)$  is slow)
- Using the Fast Walsh-Hadamard transform it is possible to speed up the sketching process ( $\Theta(d \log d)$  instead of  $\Theta(d^2)$  for matrix products)
- Empirical studies on simulated and real data show that with proper setup, learning from sketch is comparable to classical approaches (k-means clustering)

# Conclusion

Sketching is a framework that allows:

- to get leverage of assumptions on the simplicity of the underlying data generating process
- to apply ideas similar to Compressed Sensing
- learn the underlying data distribution through a compressed dataset
- not to reduce the dimension of the features, but the dimension of the dataset, without losing its characteristic (geometric) properties

