Measure compression in generative and unsupervised learning

Gabriel Turinici

CEREMADE
Université Paris Dauphine - PSL
Paris, France

Workshop "Current Trends in Applied Mathematics"
Octav Mayer Institute of Iași, Romanian Academy

Iași, November 20, 2021

Outline

Motivation: summarize information from a set of objects

Procedure at the mid-way between clustering and compression

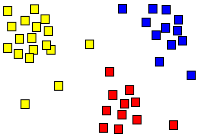


Figure: Left: example of clustering.
Middle and right: compression of the middle image into the right image (credits: Wikipedia)

Motivation: summarize information from a set of objects

Example 1: alternative to Monte Carlo to approximate integrals over high dimensional spaces : for $\int_\Omega f(\omega)d\omega$ it is good to have a sample $\frac{1}{K}\sum_{k=1}^K \delta_{\omega_k}$ close, as measure, to $d\omega$: if $d\omega \simeq \frac{1}{K}\sum_{k=1}^K \delta_{\omega_k}$ then $\int_\Omega f(\omega)d\omega \simeq \frac{1}{K}\sum_{k=1}^K f(\omega_k)$

- lower dimensional objects : quadrature;
- more exotic objects: ω (a curve) is a realization of a W_t = Brownian mvt. "cubature". $\int f(t, W_t)dW_t$

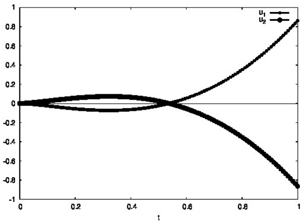


Figure: Example of cubature "points" for a Brownian motion, from [3].

Motivation: summarize information from a set of objects

Example 2: summarize a distribution with K points, e.g. 2D Gaussian.

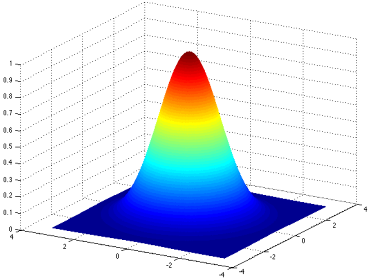


Figure: 2D Gaussian (credits: Wikipedia)

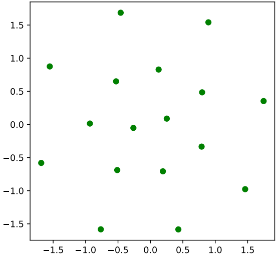


Figure: Example of compression with $K = 17$ points of a 2D Gaussian using special statistical distances (cf. [4]).
Presence of a three layers point structure: inner 2, middle 7, outer 8 (from [5]).

Motivation: summarize information from a set of objects

Example 3: summarize a large database of objects (e.g. MNIST, FMNIST, CIFAR10, ...)

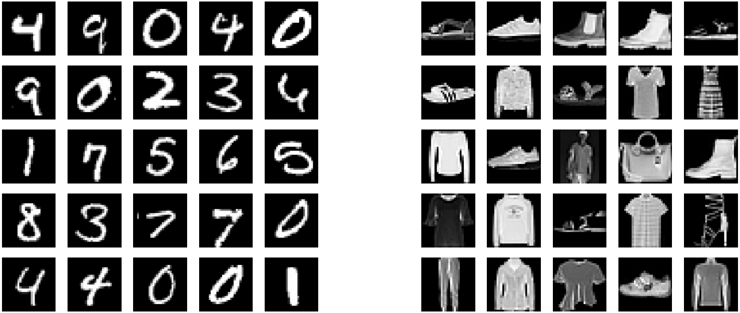


Figure: Left: MNIST samples (25 out of 60'000). Right: Fashion MNIST samples (25 out of 60'000), from [4]

Outline

Compression tools and algorithms: idea

Goal: compress a measure (usually a probability measure).
Rq: close similar to vector quantization and clustering, that aim to assign each point to a cluster.
Idea: suppose target μ is a finite Borel measure. Using the ideas from `doi:10.5281/zenodo.5705389`, to obtain a K-compression of the measure μ one minimizes the distance from $\frac{1}{K}\sum_{k=1}^{K}\delta_{x_k}$ to μ, defined as:

$$d^2\left(\frac{1}{K}\sum_{k=1}^{K}\delta_{x_k},\mu\right) := c(h,\mu) - \frac{1}{2K^2}\sum_{k\neq l}^{K} k(x_k,x_l) + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{y\sim\mu}k(x_k,y) \tag{1}$$

$k(x,y) = d(\delta_x,\delta_y)^2$; when $k(x,y) = h(|x-y|)$: translation and rotation invariant kernel statistical distance; $h(x)$ = important function to choose.

$$h = |\cdot| : \min\left(c(h,\mu) - \frac{1}{2K^2}\sum_{k\neq l}^{K}|x_k - x_l| + \frac{1}{K}\sum_{k=1}^{K} g_\mu(x_k)\right) \tag{2}$$

Question: what function h to choose ?

Definition (conditional positive definite)

A kernel $k(\cdot, \cdot)$ is said to be conditionally positive definite if for any $I \in \mathbb{N}$, $p_1, ..., p_I$ with $\sum p_i = 0$ and any $x_1, ..., x_I$: $\sum_{i,j} p_i p_j k(x_i, x_j) \geq 0$.

$-k$ is also said to be a negative definite kernel.

Theorem ("Gini difference" Gini 1912; "energy distance" Szekelly 1985, 2002; "maximum mean discrepancy" Gretton 2007, Radon-Sobolev G.T. 2021 [4])

The kernel $h(x) = |x|$ is conditionally positive definite.

Rq: many other kernels are known to be conditionally positive definite: Gaussian, etc.

Historical idea: the "energy distance" builds on the Newton's potential energy concept, cf Szekely 2002.

Statistical distances: conditionally positive kernels

Proof (GT 2021 version).

Radon transform of the dual of the homogeneous Sobolev space $\dot{H}^1$: take all directions on the unit sphere, project, measure in $\dot{H}^{-1}$, sum up: $d(\mu,\nu)^2 = \frac{1}{\text{area}(\mathbb{S})} \int_{\mathbb{S}} \|\theta_{\#}\mu - \theta_{\#}\nu\|^2_{\dot{H}^{-1}} d\theta$. Obviously positive, non-degenerate by properties of the Radon transform. □

When $d(\delta_x, \delta_y)^2 = |x - y|$, one minimizes terms involving $|\cdot|$ (not $|\cdot|^2$) : gradient descent methods experience instabilities as the differential is $\frac{x}{|x|^2}$.

Theorem (Schoenberg 1938 [2], Micchelli 1984 [1], GT 2021 [5])

For any $a \geq 0$, $\alpha \in]0,1[$, the kernels $h(x) = (a + |x|^2)^\alpha$ and $h(x) = \frac{\|x^2\|}{(a+|x|^2)^\alpha}$ are conditionally positive definite and can be expressed explicitly as a Gaussian mixture. In particular this is true for $\sqrt{a + x^2}$.

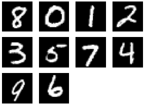
Rq: the proof extends to a larger family of kernels.

Implementation : minimize $X = (x_1, ..., x_K) \mapsto d^2\left(\frac{1}{K}\sum_{k=1}^K \delta_{x_k}, \mu\right)$

• deterministic optimization techniques when $x \mapsto \mathbb{E}_{y\sim\mu} h(x-y)$ has a closed form (e.g. normal mixture)

• ML / stochastic optimization algorithms (e.g. SGD, Adam, momentum, ...) when the database is large: compute a noisy gradient using batches/sampling from the database.

Good convergence is obtained in general.



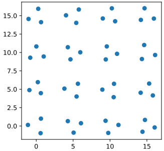


Figure: Measure compression results, from [5].
Left : MNIST compression with $K = 10$ samples: we computed the compression then took closest from the database.Note that algorithm chooses by itself to represent the each figure exactly once. **Right :** 16 2D Gaussians on a grid compressed with $K = 16 * 3$ points.

Outline

Theoretical questions

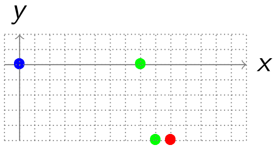
• does the minimization of $X = (x_1, ..., x_K) \mapsto d^2 \left(\frac{1}{K} \sum_{k=1}^{K} \delta_{x_k}, \mu \right)$ has a solution ? Yes (standard continuity and compactness).

• existence for non-uniform compression weights ? OK, minimize, for given p_k that sum to 1 : $X = (x_1, ..., x_K) \mapsto d^2 \left(\sum_{k=1}^{K} p_k \delta_{x_k}, \mu \right)$.

• what about p being also a variable (clusters of unknown weight) ? OK, the optimization w/r to p is analytic.

Theoretical questions: positivity of the compression

Question: suppose $\mu \geq 0$; when optimizing both weights p_k and support points x_k variables) is the compressed measure positive too ?



projection of $(0,0)$ to $(8,0),(9,-5),(10,-5)$ has optimal weight negative for red point

Proposition (..., GT 2021 in some cases)

Let μ be a probability law on a convex domain with finite first order moment and $K \in \mathbb{N}^\star$. If

$$\sum_{k=1}^{K} p_k^* \delta_{x_k^*} \in argmin_{p_k, x_k, \sum_{k=1}^{K} p_k = 1} \left[d\left(\sum_{k=1}^{K} p_k \delta_{x_k}, \mu \right)^2 \right] \tag{3}$$

then $\sum_{k=1}^{K} p_k^* \delta_{x_k^*} \geq 0$ *i.e.,* $p_k^* \geq 0, \ \forall k$.

Theoretical questions: non-constant target compression

• Question: how does the compression depends on target measure μ ? Suppose μ depends on parameter u e.g. $\mu(u) = \mathcal{N}(u, 1)$ (1D normal of mean u, variance 1). K = fixed, compression for given u =ok. What about continuity w/r to u ?

• $\mu = \mu(u)$, each measure is 1D valued; to simplify we take them as probability laws.

Lemma (regularity w/r to target)

Suppose $u \mapsto \mu(u)$ is regular enough (...) and the measure $\mu(u)$ is non-atomic $\forall u$; then:

- *the minimization problem $\frac{1}{K}\sum_{k=1}^{K}\delta_{x_k} \mapsto d^2\left(\frac{1}{K}\sum_{k=1}^{K}\delta_{x_k}, \mu(u)\right)$ admits a unique solution $C(u) = \frac{1}{K}\sum_{k=1}^{K}\delta_{x_k}$ (as a probability law);*
- *the mapping $u \mapsto C(u)$ is regular with respect to u.*

Theoretical questions: non-constant target compression

Multi-D : $u \in \Omega \mapsto \mu(u)$, measure on $\mathbb{R}^N$

Problems :

• the compression is not necessarily unique for some u (e.g. symmetries of μ);

• difficult to prove the existence of a continuous selection (e.g. Kakutani et al.) ... lack of convexity.

Example: minimize norm of $C : \Omega \to \mathbb{R}^{K \times N}$ in some Sobolev space H

$$\varepsilon \| C(\cdot) \|_H^2 + \int_\Omega d\left(\frac{1}{K} \sum_{k=1}^{K} \delta_{C(u)_k}, \mu(u) \right)^2 du \tag{4}$$

$$= \varepsilon \| C(\cdot) \|_H^2 + \int_\Omega \mathcal{F}(u, C(u)) du \tag{5}$$

Remark: existence ok, but uniformity when $\varepsilon \to 0$ not clear.

Conclusions and future work

Further questions:

- more details on the topology depending on h
- how to interpolate, continuous selection ?
- positivity under more general conditions

References

[1] Charles A Micchelli. "Interpolation of scattered data: distance matrices and conditionally positive definite functions". In: *Approximation theory and spline functions*. Springer, 1984, pp. 143–145.

[2] Isaac J Schoenberg. "Metric spaces and completely monotone functions". In: *Annals of Mathematics* (1938), pp. 811–841.

[3] Gabriel Turinici. "Cubature on C1 Space". In: *Control and Optimization with PDE Constraints*. Springer, 2013, pp. 159–172.

[4] Gabriel Turinici. "Radon–Sobolev Variational Auto-Encoders". In: *Neural Networks* 141 (2021), pp. 294–305. ISSN: 0893-6080. DOI: `10.1016/j.neunet.2021.04.018`.

[5] Gabriel Turinici. "Unbiased metric measure compression". 2021. DOI: `10.5281/zenodo.5705389`.