# Algorithms that get old : the case of generative deep neural networks

Gabriel Turinici

CEREMADE
Université Paris Dauphine - PSL
Paris, France

LOD 2022,
The 8th International Conference on
Machine Learning, Optimization, and Data Science,

Italy, September 18 – 22, 2022

# Outline

1. Introduction

2. Technical ingredients
   - Statistical distances and conditionally negative kernels

3. Numerical results
   - Diverse but history unaware sampling
     - Diverse and history aware multi-D Gaussian sampling and application to generative algorithms

# Introduction and motivation

• we are concerned with GENERATIVE algorithms i.e. that create new objects (e.g., images) based on some database

• We want to avoid repetitions and enforce diversity in this creation, like human painters do not paint twice same painting, have "periods", same for writers, musicians, ...

Famous painters have "periods" : here Pablo Picasso's rose, blue, cubism, surrealism periods.

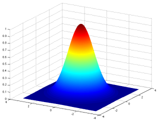
(from https://mymodernmet.com/pablo-picasso-periods/)

# Introduction and motivation: mathematical framework

• Given : empirical database $\mu_e = \frac{1}{M}\sum_{\ell=1}^{M} \delta_{x_\ell}$ sampling from unknown distribution $\mu$ ($x_\ell \sim \mu$).

• Goal: construct samples as $\mu$

Example: sampling from 2D Gaussian distribution results in most samples in the red part.



• Problem : samples are often not so diverse; example for a GAN / VAE, sampling is done from the latent distribution with replacement.

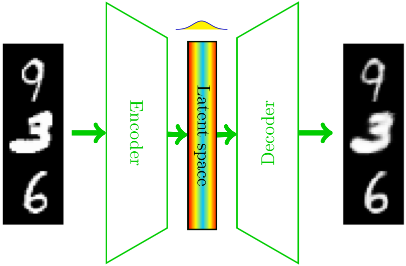
• Idea: make the algorithm keep the memory of previous actions and thus instaurate irreversibility (and dependence of the past) in the choices, that we call "age".

# Introduction and motivation: technical framework

• Idea: make the algorithm keep the memory of previous actions and thus instaurate irreversibility (and dependence of the past) in the choices, that we call "age".

• Question: what is $\mu$ in practice ?

Variational Autoencoder (VAE) structure: the available data is used to train two networks (encoder and decoder) to reproduce it and obtain a reference distribution (here in yellow) on the latent space; then in the generation phase the decoder is used to create new data.



# Outline

1 Introduction

2 Technical ingredients

- Statistical distances and conditionally negative kernels

3 Numerical results

- Diverse but history unaware sampling
  - Diverse and history aware multi-D Gaussian sampling and application to generative algorithms

# Technical ingredients

Goal: incremental procedure
- find $K$ new samples ($K$ Dirac masses centered at some $x_k$, $k = 1, ..., K$) from the target measure $\mu$
- takes into account the historical points points $Y = (y_j)_{j=1}^{K_p}$ (already available): "only add what is missing"

## Mathematical formulation

Find the multi-point $X = (x_k)_{k=1}^{K} \in \mathbb{R}^{N \times K}$ ($k = 1, ..., K$) that minimizes the distance from the total empirical distribution $\frac{\sum_{k=1}^{K_p} \delta_{y_k} + \sum_{l=1}^{K} \delta_{x_k}}{K_p + K}$ to the target measure $\mu$ ($y_k$ are given).

$$\text{Equivalent formulation : minimize } X \mapsto dist(\delta_X, \eta)^2, \tag{1}$$

where $\delta_X := \frac{1}{K} \sum_{l=1}^{K} \delta_{x_k}$, $\delta_Y = \frac{1}{K_p} \sum_{k}^{K_p} \delta_{y_k}$, $\eta = \frac{(K_p+K)\mu - K_p \delta_Y}{K_P + K}$.
Remarks: $Y$ is given (previous choices), $\eta$ is a signed measure !

# Technical how-to

Questions:

- how to compute $X \mapsto dist(\delta_X, \eta)^2$
- how to minimize it ?

Distance: use a conditionally negative definite kernel $h$ :

$$d(\eta_1, \eta_2)^2 = \int\int h(|X - Y|)(\eta_1 - \eta_2)(dX)(\eta_1 - \eta_2)(dY). \tag{2}$$

In particuler for discrete distributions $\eta_i = \sum_{k=1}^{K_i} p_k^i \delta_{z_k^i}$ :

$$d(\eta_1, \eta_2)^2 = \sum_{k=1}^{K_1} \sum_{\ell=1}^{K_2} p_k^1 p_\ell^2 h(|z_k^1 - z_\ell^2|). \tag{3}$$

# Statistical distances: conditionally negative kernels

Question: what function $h$ to choose ?

**Definition (conditional negative definite)**

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

$-h$ is also said to be a (conditionally) positive 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 negative definite.*

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

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

# Statistical distances: conditionally negative 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 $h(x) = (a^2 + |x|^2)^\alpha - a^{2\alpha}$ and $h(x) = \frac{\|x^2\|}{(a^2+|x|^2)^\alpha}$ are conditionally negative definite and can be expressed explicitly as a Gaussian mixture. In particular this is true for $\sqrt{a^2+x^2} - a$.*

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

# Theoretical existence results

## Proposition

*Suppose $K$ is a fixed positive integer. Let $\eta$ be a signed measure such that $\int(1+|X|)\eta(dX) < \infty$ . For any vector $Z = (z_j)_{j=1}^J \in \mathbb{R}^{N\times J}$ denote*

$$\delta_Z := \frac{1}{J}\sum_{j=1}^{J}\delta_{z_j}, \;\; f(Z) := dist_{h=|\cdot|}\left(\delta_Z, \eta\right)^2 . \tag{4}$$

*Then the minimization problem :*

$$\inf_{X=(x_k)_{k=1}^K \in \mathbb{R}^{N\times K}} f(X) \tag{5}$$

*admits at least one solution.*

# Stochastic minimization algorithm

## History aware (signed measure) compression algorithm : HAW-C

- set batch size $B$, parameter $a = 10^{-6}$,
- load the historical points $y_k$, $k = 1, ..., K_p$
- initialize points $x_k$, $k = 1, ..., K$ sampled at random from $\mu$, denote $X = (x_k)_{k=1}^K$ (considered as vector in $\mathbb{R}^{N \times K}$)
- while max iteration not reached
  - sample $z_1, ..., z_B \sim \mu$ (i.i.d).
  - compute the global loss [a] using formula (3) :
    $$L(X) := d\left(\frac{1}{K}\sum_{l=1}^K \delta_{x_k}, \frac{K_p+1}{B}\sum_{b=1}^B \delta_{z_b} - \sum_{j=1}^{K_p} \delta_{y_j}\right)^2;$$
  - backpropagate the loss $L(X)$ in order to minimize $L(X)$ and update $X$.

[a] The global loss = distance from $\delta_X = \frac{1}{K}\sum_{l=1}^K \delta_{x_k}$ and $\eta$.

- deterministic optimization 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.

# Outline

1. Introduction

2. Technical ingredients
   - Statistical distances and conditionally negative kernels

3. Numerical results
   - Diverse but history unaware sampling
     - Diverse and history aware multi-D Gaussian sampling and application to generative algorithms

# Diverse but history unaware sampling of a 2D Gaussian

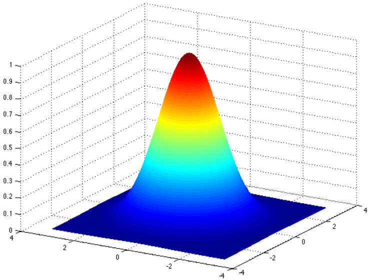


Figure: 2D Gaussian (credits: Wikipedia)

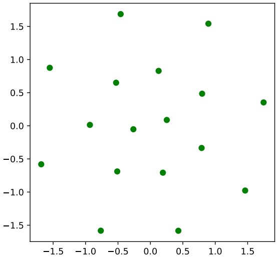
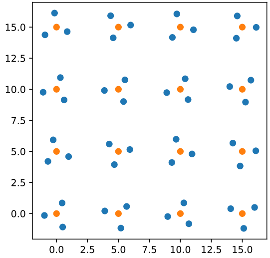


Figure: Example of compression with $K = 17$ points of a 2D Gaussian using special statistical distances (cf. [4]).
Presence of a three layers point structure: inner 2, middle 7, outer 8 (from [5]).

# Diverse but history unaware sampling of a 2D Gaussian mix distribution



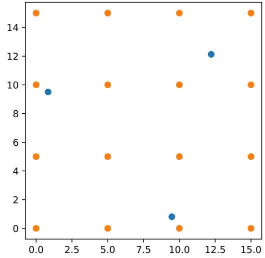
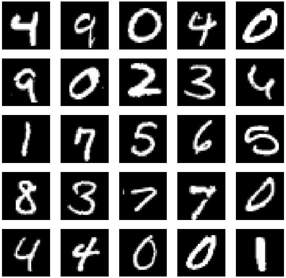




Figure: Test without any historical points, $K_p = 0$. An example of compression for an uniform Gaussian mixture of 16 Gaussians centered on points of a $4 \times 4$ grid (red points are the centers of the Gaussians, blue points are the compressed points). We used $K$ points to summarize the distribution : $K = 48$ (**left image**) or $K = 3$ (**right image**). Good quality results are obtained as the algorithm "understands" the mixing structure: for instance for $K = 48$ the algorithm allocates precisely 3 points per Gaussian mixture term.

# Application: diverse sampling from a large database (here MNIST, FMNIST)

Compression of a multi-D Gaussian is used to sample from the latent space.



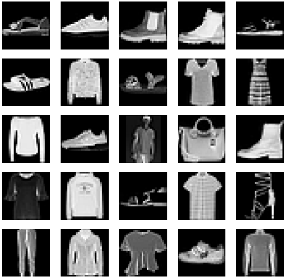


Figure: Left: MNIST samples (25 out of 60'000). Right: Fashion MNIST samples (25 out of 60'000), from [4]

# History aware multi-dimensional Gaussian compression

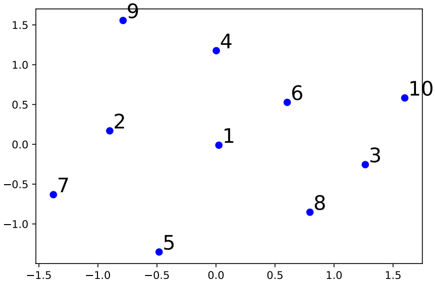
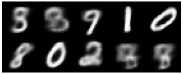
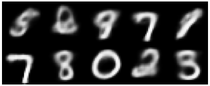


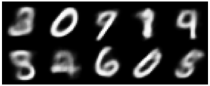

Figure: An example of diverse and history aware (recursive compression) of a $2D$ standard Gaussian; the result of the compression after 10 iterations. Each point $u_i$ is labeled by its corresponding index $i$ when it was found.

# Multi-dimensional Gaussian compression for "old" generative algorithms



Images using standard cVAE (cf. Tensorow documentation) obtained by taking either a random sampling of 10 points from a 2D Gaussian (top image) or the sampling obtained in figure 5 (bottom image). The bottom image appears more faithful of the database.





Results of the same procedure on an improved network (512 filters / 20 epochs): **top image :** random sampling; **bottom image :** decoding of the incremental sampling. The top image has several repetitions (for instance figure 7) that are absent from the bottom figure but more importantly, some figures abundant in the database and not present in the top figure appear in the other one, like the figures 1 and 6.

# 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.