Block 3: Approximation and Interpolation

Domènec Ruiz-Balet

29th August 2025

CEREMADE, Paris Dauphine
Universitat de Barcelona
Strasbourg 2025

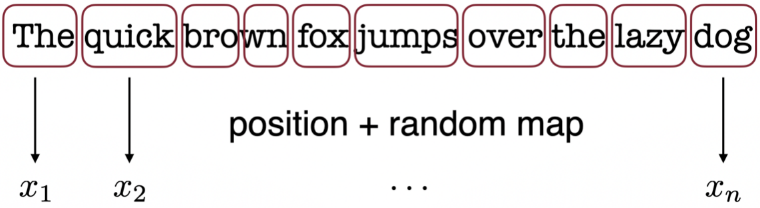
Recall on transformers

Example

The quick brown fox jumps over the lazy dog

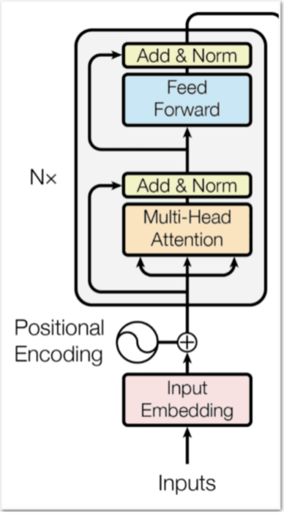
The phrase "The quick brown fox jumps over the lazy dog" is commonly used because it contains every letter of the English alphabet at least once.

The data



- Each sentence is mapped to a sequence
- d of the order of hundreds
- n of the order of the length of a paragraph, a book etc
- One can also use it in images

The Transformer



Chat GPT2 code

gpt-2 / src / model.py

Code | Blame | 174 lines (144 loc) · 6.35 KB

```python
def block(x, scope, *, past, hparams):
    with tf.variable_scope(scope):
        nx = x.shape[-1].value
        a, present = attn(norm(x, 'ln_1'), 'attn', nx, past=past, hparams=hparams)
        x = x + a
        m = mlp(norm(x, 'ln_2'), 'mlp', nx*4, hparams=hparams)
        x = x + m
        return x, present
```

https://github.com/openai/gpt-2/blob/master/src/model.py

[Radford et al. **Language Models are Unsupervised Multitask Learners**]

Chat GPT2

Recall that we had $(x_1, ..., x_n) \in \mathbb{R}^{nd}$.

- **Attention**

$$\mathbf{V}^t \sum_{j=1}^{n} \frac{e^{\langle \mathbf{B}^t x_i, x_j \rangle} x_j}{\sum_{k=1}^{n} e^{\langle \mathbf{B}^t x_i, x_k \rangle}}$$

- **MLP**

$$\mathbf{W}^t \sigma(\mathbf{U}^t x + b^t)$$

- We have a controlled discrete dynamical system
- $\mathbf{V}^t, \mathbf{B}^t, \mathbf{W}^t, \mathbf{U}^t, b^t$ are parameters to be chosen

$$y_i^t = x_i^t + \mathbf{V}^t \sum_{j=1}^{n} \frac{e^{\langle \mathbf{B}^t x_i, x_j \rangle} x_j}{\sum_{k=1}^{n} e^{\langle \mathbf{B}^t x_i, x_k \rangle}}$$

$$x_i^{t+1} = \frac{y_i^t + \mathbf{W}^t \sigma(\mathbf{U}^t y_i^t + b^t)}{\|\mathbf{W}^t \sigma(\mathbf{U}^t y_i^t + b^t)\|_2}$$

Can we obtain a differential equation?

$$y_i^t = x_i^t + \mathbf{V}^t \sum_{j=1}^{n} \frac{e^{\langle \mathbf{B}^t x_i, x_j \rangle} x_j}{\sum_{k=1}^{n} e^{\langle \mathbf{B}^t x_i, x_k \rangle}} \qquad x_i^{t+1} = \frac{y_i^t + \mathbf{W}^t \sigma(\mathbf{U}^t y_i^t + b^t)}{\|\mathbf{W}^t \sigma(\mathbf{U}^t y_i^t + b^t)\|_2}$$

- Set $\mathbf{V}^t = \Delta t \tilde{\mathbf{V}}^t$ and $\mathbf{W}^t = \Delta t \tilde{\mathbf{W}}^t$ + Taylor + $\Delta t \to 0$
- Lie-Trotter splitting of an ODE!

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\boldsymbol{V}(t) \sum_{j=1}^{n} \frac{e^{\langle \mathbf{B}(t) x_i(t), x_j(t) \rangle} x_j(t)}{\sum_{k=1}^{n} e^{\langle \mathbf{B}(t) x_i(t), x_k(t) \rangle}} + \boldsymbol{W}(t) \sigma\big(\boldsymbol{U}(t) x_i(t) + b(t)\big) \right)$$

where

$$\mathbf{P}^{\perp}_{x}(v) = v - \langle v, x \rangle x$$

[Lu et al. 2019]
[Geshkovski, Letrouit,Polyanskiy, Rigollet 2023]

[Collective behaviour literature: Carrillo, Ha, Tadmor, Trélat, ...]

Differential equation

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\boldsymbol{V}(t) \sum_{j=1}^{n} \frac{e^{\langle \mathbf{B}(t) x_i(t), x_j(t) \rangle}}{\sum_{k=1}^{n} e^{\langle \mathbf{B}(t) x_i(t), x_k(t) \rangle}} x_j(t) + \boldsymbol{W}(t) \sigma\big(\boldsymbol{U}(t) x_i(t) + b(t) \big) \right)$$

- $\boldsymbol{W}(t) = 0$ Self-attention dynamics

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\boldsymbol{V}(t) \sum_{j=1}^{n} \frac{e^{\langle \mathbf{B}(t) x_i(t), x_j(t) \rangle}}{\sum_{k=1}^{n} e^{\langle \mathbf{B}(t) x_i(t), x_k(t) \rangle}} x_j(t) \right)$$

- $\boldsymbol{V}(t) = 0$ Neural ODE on the Sphere

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \big(\boldsymbol{W}(t) \sigma\big(\boldsymbol{U}(t) x_i(t) + b(t) \big) \big)$$

Neural ODEs

Residual Neural Networks, for $t = 0, ..., T$:

$$\begin{cases} x(t+1) = x(t) + \Delta t \boldsymbol{W}(t)\sigma(\boldsymbol{U}(t)x(t) + b(t)) \\ x(0) = x_0 \end{cases}$$

- $\Delta t \to 0$. Connection via discretizations of the Neural ODE

$$\begin{cases} \dot{x}(t) = \boldsymbol{W}(t)\sigma(\boldsymbol{U}(t)x + b(t)) \\ x(0) = x_0 \end{cases}$$

- Representation with infinitely many layers
- Now the parameters are functions of time
  $\boldsymbol{W}, \boldsymbol{U} \in L^\infty((0,T); \mathbb{R}^{d\times d})$ and $b \in L^\infty((0,T); \mathbb{R}^{d\times d})$

[Weinan E 2017]

[Chen et al 2018]

Neural ODEs

Neural ODEs are parameterized flow maps in $\mathbb{R}^d$.

$$\begin{cases} \dot{x}(t) = \boldsymbol{W}(t)\sigma(\boldsymbol{U}(t)x + b(t)) \\ x(0) = x_0 \end{cases}$$

The solution map is a (parameterized) function

$$f_\theta{}^T : \mathbb{R}^d \to \mathbb{R}^d$$

$$f_\theta{}^T(x_0) = x(T)$$

Flow maps

- Neural ODEs are parameterized flow maps in $\mathbb{R}^d$

[Li, Lin, Shen 2022]
[**R-B**,Zuazua 2023]
[Cheng, Li, Lin, Shen 2024]

The problem

The input of the transformer

The size of the sequence

- Each input may have a different length
- n can be very large, like a paragraph or a book.

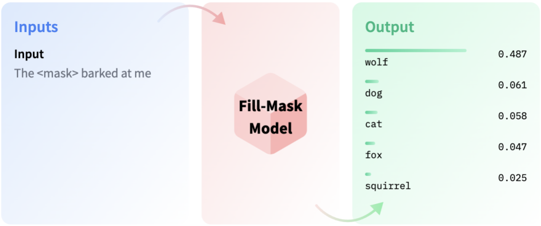
Modeling it via measures can be the right setting

$$(x_1, x_2, ..., x_n) \longrightarrow \frac{1}{n}\sum_{j=1}^{n}\delta_{x_j}$$

We could even think that is an AC measure

The output of the transformer

Masked language modeling

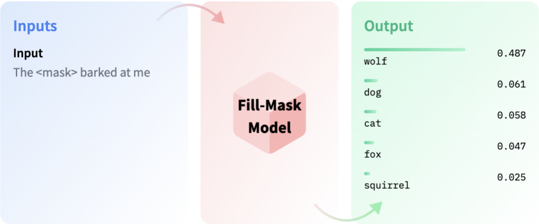


The output measure has very few atoms

$$\mu_1 = \sum_{j=1}^{m} \alpha_j^i \delta_{z_j} \qquad\qquad n \gg m$$

Practical objective

Masked language modeling



Q1: Can we match N inputs to N outputs with a Transformer?

$$\mu_0^i = \frac{1}{n}\sum_{j=1}^{n} \delta_{x_j^i} \longrightarrow \mu_1^i = \sum_{j=1}^{m} \alpha_j^i \delta_{z_j^i} \qquad i = 1, \ldots, N$$

The transformer for general measures

Since the order does not matter

$$(x_1, x_2, ..., x_n) \longrightarrow \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$$

And more generally we can write a parameterized mean-field equation

$$\partial_t \mu + \operatorname{div}\left(v_\theta[\mu(t)](x,t)\mu\right) = 0$$

where

$$v_\theta[\mu](x,t) = \mathbf{P}_x^{\perp}\left(\boldsymbol{V}(t)\mathscr{A}_{\boldsymbol{B}(t)}[\mu(t)] + \boldsymbol{W}(t)\sigma\left(\boldsymbol{U}(t)x_i(t) + b(t)\right)\right)$$

$$\mathscr{A}_{\boldsymbol{B}(t)}[\mu(t)] = \int \frac{e^{\langle \mathbf{B}(t)x, y\rangle} y \mu(\mathrm{d}y, t)}{\int e^{\langle \mathbf{B}(t)x, z\rangle}\mu(\mathrm{d}z, t)}$$

Flow maps in $\mathscr{P}(\mathbb{S}^{d-1})$

Since the Cauchy problem

$$\begin{cases} \partial_t \mu + \operatorname{div}\left(v_\theta[\mu(t)](x,t)\mu\right) = 0 \\ \mu(0) = \mu_0 \end{cases}$$

is well posed, we have a parameterized flow map in $\mathscr{P}(\mathbb{S}^{d-1})$.
The solution map is a (parameterized) function

$$\Phi_\theta{}^T : \mathscr{P}(\mathbb{S}^{d-1}) \to \mathscr{P}(\mathbb{S}^{d-1})$$

$$\Phi_\theta{}^T(\mu_0) = \mu(T)$$

Flow maps

- Neural ODEs are parameterized flow maps in $\mathbb{R}^d$
- Transformers are parameterized flow maps in $\mathscr{P}(\mathbb{S}^{d-1})$

[Sander, Albin, Blondel, Peyré 2022]

[Geshkovski, Letrouit,Polyanskiy, Rigollet 2023]

Matching ensembles of measures

Q1: Can we match N inputs (discrete) to N outputs (discrete) with a Transformer?

This is a question in transport theory

Given initial and target probability measures

$$\{\mu_0^k\}_{k\in[N]} \subset \mathscr{P}(\mathbb{S}^{d-1}) \qquad \{\mu_T^k\}_{k\in[N]} \subset \mathscr{P}(\mathbb{S}^{d-1})$$

Interpolation problem (informal)

Can we find $\Phi_\theta^T : \mathscr{P}(\mathbb{S}^{d-1}) \to \mathscr{P}(\mathbb{S}^{d-1})$ s.t.$\Phi_\theta^T(\mu_0^k) \approx \mu_T^k \qquad k = 1, ..., N$

By lifting to general measures, we address large n

The map Φ_θ^T is the same for all k!

In other words, θ is the same for all $k \implies$ Simultaneous/ensemble control

Universal approximation: Furuya-de Hoop-Peyré 2024

Furuya-de Hoop-Peyré theorem

$$\Gamma_\theta(\mu, x) := x + \sum_{h=1}^{H} W^h \int \frac{\exp\left(\frac{1}{\sqrt{k}}\langle Q^h x,\, K^h y\rangle\right)}{\int \exp\left(\frac{1}{\sqrt{k}}\langle Q^h x,\, K^h z\rangle\right) \mathrm{d}\mu(z)} V^h y \, \mathrm{d}\mu(y).$$

$$(\Gamma_2 \diamond \Gamma_1)(\mu, x) := \Gamma_2(\mu_1, \Gamma_1(\mu, x)), \quad \text{where} \quad \mu_1 := \Gamma_1(\mu)_\sharp \mu,$$

Theorem 1. *Let $\Omega \subset \mathbb{R}^d$ be a compact set and $\Lambda^\star : \mathcal{P}(\Omega) \times \Omega \to \mathbb{R}^{d'}$ be continuous, where $\mathcal{P}(\Omega)$ is endowed with the weak* topology. Then for all $\varepsilon > 0$, there exist L and parameters $(\theta_\ell, \xi_\ell)_{\ell=1}^L$, such that*

$$\forall (\mu, x) \in \mathcal{P}(\Omega) \times \Omega, \quad |F_{\xi_L} \diamond \Gamma_{\theta_L} \diamond \ldots \diamond F_{\xi_1} \diamond \Gamma_{\theta_1}(\mu, x) - \Lambda^\star(\mu, x)| \leq \varepsilon,$$

with $d_{\text{in}}(\theta_\ell) \leq d + 3d'$, $d_{\text{head}}(\theta_\ell) = k(\theta_\ell) = 1$, $H(\theta_\ell) \leq d'$.

1. Proof by Stone-Weierstrass
2. No quantification, everything implicit
3. Also results for masked attention models

Why the non-linearity?

Why not to consider a linear continuity equation

$$\begin{cases} \partial_t \mu + \mathrm{div}(V(x,t)\mu) = 0 & (x,t) \in \mathbb{S}^{d-1} \times (0,1) \\ \mu(0) = \mu_0 \end{cases}$$

The solution can be written as

$$\mu(1) = \mathrm{T}_{\#}\mu_0$$

where the map $\mathrm{T} : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ is computed solving the ODE

$$\begin{cases} \dot{y}(t;x) = V(y(t;x),t) & t \in (0,1) \\ y(0;x) = x \end{cases}$$

and then

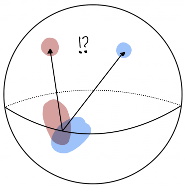
$$\mathrm{T}(x) = y(1;x)$$

Why the non-linearity?

Assume that

$$\mathrm{supp}(\mu_0^1) \cap \mathrm{supp}(\mu_0^2) \neq \emptyset \qquad \mathrm{supp}(\mu_T^1) \cap \mathrm{supp}(\mu_T^2) = \emptyset$$

Then, the problem is unfeasible, since the transport map would have to be multivalued in $\mathrm{supp}(\mu_0^1) \cap \mathrm{supp}(\mu_0^2)$!



Ensemble matching cannot be done with a linear continuity equation!

BUT TRANSFORMERS ARE NONLINEAR IN μ!!

Results

Theorem

Theorem

Suppose $d \geq 3$. *Assume that*
For any $1 \leq i \leq N$, *there exists* $\mathrm{T}^i \in L^2(\mathbb{S}^{d-1}; \mathbb{S}^{d-1})$ *such that* $\mathrm{T}^i_{\#}\mu^i_0 = \mu^i_1$.
Then for any $T > 0$ *and* $\varepsilon > 0$, $\exists\theta$ *piece-wise constant s.t. for any* $1 \leq i \leq N$, *the solution* $\mu^i \in \mathscr{C}^0([0, T]; \mathscr{P}(\mathbb{S}^{d-1}))$ *satisfies*

$$\mathrm{W}_2\left(\mu^i(T), \mu^i_1\right) \leq \varepsilon.$$

On the assumption

The assumption is minimal! We cannot split a Dirac mass in two

$$\mu_0 = \delta_{x_0}, \qquad \mu_1 = \frac{1}{2}\delta_{x_1} + \frac{1}{2}\delta_{x_2}$$

(the vector field of the transformer is Lipschitz)

A simpler and quantified statement

Theorem

Suppose $d \geq 3$ and $m \in \mathbb{N}$ and consider for every $i = 1, ..., N$

$$\mu_0^i = \frac{1}{n}\sum_{j=1}^{n} \delta_{x_j} \qquad \text{(or } \mu_0^i \text{ AC)} \qquad \text{and} \qquad \mu_1^i = \frac{1}{m}\sum_{j=1}^{m} \delta_{z_j}$$

with $n \gg m$ and n multiple of m. Then for any $T > 0$ and $\varepsilon > 0$, $\exists \theta$ piece-wise constant s.t. for any $1 \leq i \leq N$, the solution satisfies

$$W_2\left(\mu^i(T), \mu_1^i\right) \leq \varepsilon.$$

$$\#\mathrm{disc}(\theta) = O(mN) \qquad \text{(independent of } n\text{!)}$$

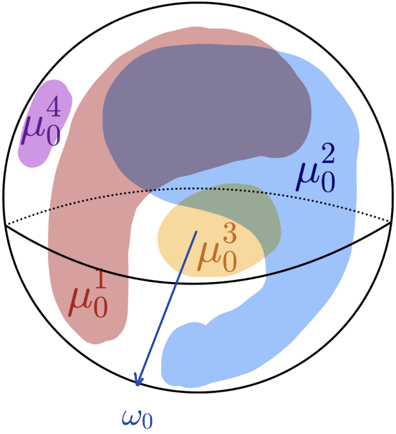
Estimating the number of parameters is key

Hints on the proofs

Technical (removable) assumption

Assume that there exist $\omega_0, \omega_1 \in \mathbb{S}^{d-1}$ such that

$$\omega_0 \notin \bigcup_{1\leq i\leq N} \text{supp}(\mu_0^i) \quad \text{and} \quad \omega_1 \notin \bigcup_{1\leq i\leq N} \text{supp}(\mu_1^i)$$



Idea of the proof

Split the time interval in three

$$[0,T] = \left[0, \frac{T}{3}\right] \cup \left[\frac{T}{3}, \frac{2T}{3}\right] \cup \left[\frac{2T}{3}, T\right]$$

and we seek to build three flow maps

$$\Phi^t_{\theta_1}, \Phi^t_{\theta_2}, \Phi^t_{\theta_3} : \mathscr{P}(\mathbb{S}^{d-1}) \to \mathscr{P}(\mathbb{S}^{d-1})$$

such that

1. $\Phi^t_{\theta_1}$ and $\Phi^t_{\theta_3}$ are such

$$\text{supp}(\Phi^T_{\theta_1}(\mu^i_0)) \cap \text{supp}(\Phi^T_{\theta_1}(\mu^j_0)) = \emptyset$$

$$\text{supp}(\Phi^T_{\theta_3}(\mu^i_1)) \cap \text{supp}(\Phi^T_{\theta_3}(\mu^j_1)) = \emptyset$$

   (Disentangle supports)

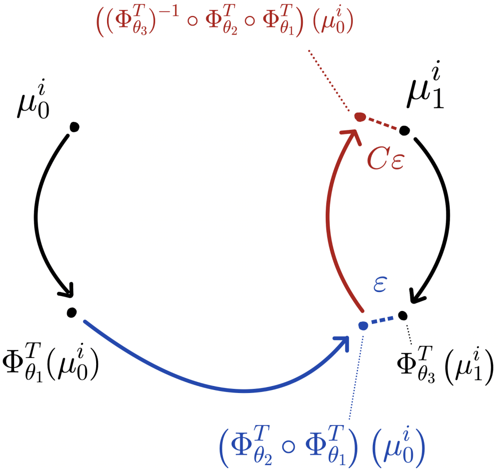
2. $\Phi^t_{\theta_2}$ can approximately match disentangled measures

$$W_2\left((\Phi^T_{\theta_2} \circ \Phi^T_{\theta_1})(\mu^i_0), \Phi^T_{\theta_3}(\mu^i_1)\right) \leq \varepsilon$$

   (Matching disentangled measures)

Idea of the proof

1. $\Phi^t_{\theta_1}$ and $\Phi^t_{\theta_3}$ (Disentangle supports)
2. $\Phi^t_{\theta_2}$ (Matching disentangled measures)



Idea of the proof

Then the solution at the final state

$$\mu(T)^i = \left((\Phi^T_{\theta_3})^{-1} \circ \Phi^T_{\theta_2} \circ \Phi^T_{\theta_1} \right) (\mu^i_0)$$

satisfies that

$$\begin{aligned} W_2\left(\mu^i(T), \mu^i_1\right) &= W_2\left(\left((\Phi^T_{\theta_3})^{-1} \circ \Phi^T_{\theta_2} \circ \Phi^T_{\theta_1} \right) (\mu^i_0), \left((\Phi^T_{\theta_3})^{-1} \circ \Phi^T_{\theta_3} \right) (\mu^i_1) \right) \\ &\lesssim_{T,\theta_3} W_2\left((\Phi^T_{\theta_2} \circ \Phi^T_{\theta_1}) (\mu^i_0), (\Phi^T_{\theta_3})(\mu^i_1) \right) \\ &\lesssim_{T,\theta_3} \varepsilon. \end{aligned}$$

- The Lipschitz vector field, and the continuous time, gives us automatically that the flow map is invertible.
- Lie-Bracket analogy in nonlinear control
- Remains to build the maps $\Phi^T_{\theta_1}, \Phi^T_{\theta_2}, \Phi^T_{\theta_3}$
- Disentangle supports via Self-Attention dynamics
- Matching disentangled measures via Neural ODEs

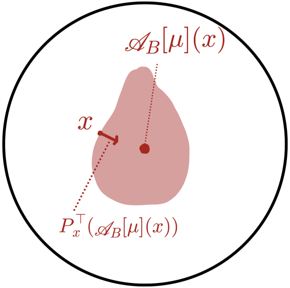
Clustering

Clustering

Set $\boldsymbol{W} = 0$,
If $\text{supp}(\mu_0)$ is contained in Half-sphere.
Then the solution with $\boldsymbol{V} = I_d$ (Self-attention) satisfies

$$\lim_{t \to +\infty} \mathrm{W}_2(\mu(t), \delta_z) = 0$$

for some $z \in \text{supp}(\mu_0)$.



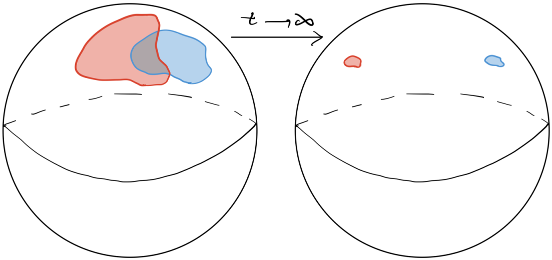
Disentanglement

Set $\boldsymbol{W} = 0$

If we had clustering with known limit locations, e.g. $\boldsymbol{V} = I_d$, $\boldsymbol{B} = \beta I_d$

$$\lim_{t \to +\infty} \mathrm{W}_2(\mu^i(t), \delta_{z_i})$$

for some $z^i \neq z^j$ we would be done. But we don't in general.



The average is all you need

Set $N = 2$. We consider $\mu_0^1, \mu_0^2 \in \mathscr{P}(\mathbb{Q}_1)$ and

$$\partial_t \mu(t) + \mathrm{div}(P_x V(t) \mathbb{E}_{\mu(t)}[x]\mu(t)) = 0$$

Suppose $\mathbb{E}_{\mu_0^1}[x]$ is not colinear with $\mathbb{E}_{\mu_0^2}[x]$.
Take

$$\boldsymbol{V}(t) = \sum_{j=1}^{d-1} \alpha_j \alpha_j^\top 1_{[T_j, T_{j+1}]}(t)$$

with α_j's being an orthonormal basis of

$$(\mathrm{span}\mathbb{E}_{\mu_0^1}[z])^\top$$

Then, there is some j for which

$$\left\langle \mathbb{E}_{\mu_0^1}[z], \alpha_j \right\rangle = 0, \qquad \left\langle \mathbb{E}_{\mu_0^2}[z], \alpha_j \right\rangle \neq 0$$

The average is all you need

Then compute

$$\frac{d}{dt}\langle \mathbb{E}_{\mu^i(t)[z]}, \alpha_j\rangle = \langle \mathbb{E}_{\mu^i(t)[z]}, \alpha_j\rangle \left(1 - \int \langle y, \alpha_j\rangle^2 \mu^i(t, \mathrm{d}y)\right)$$

1. $\langle \mathbb{E}_{\mu^i(t)[z]}, \alpha_j\rangle$ does not change sign
2. $\langle \mathbb{E}_{\mu^i(t)[z]}, \alpha_j\rangle = 0$ for $i = 1$ and for $\mu = \delta_{\pm\alpha_j}$

Therefore for any $x(t) \in \mathrm{supp}(\mu^i(t))$ we have that

$$\frac{d}{dt}\langle x(t), \alpha_j\rangle = \left\langle \alpha_j, \mathbb{E}_{\mu(t)}[x]\right\rangle \left(1 - \langle \alpha_j, x(t)\rangle^2\right)$$

1. $\mu^1(t) = \mu_0^1$ for all $t \geq 0$
2. μ_0^2 moves to $\pm\alpha_j$ ($\alpha_j \notin \mathbb{Q}_1$)

Matching disentangled measures

The key ingredients are

- Quantize the target (disentangled) measure (if needed).
- Cluster the input measure in m atoms
- Interpolation (simultaneous control) of $M = mN$ to $M = mN$ points in $\mathbb{S}^{d-1}$.

Clustering

Clustering will allow to reduce the problem to an interpolation problem

Estimates

The strategy gives a straightforward way to estimate the number of discontinuities

$$\#Disc_{Total} = \#Disc_{Sep.} + \#Disc_{Clus.} + \#Disc_{Interp.}$$

Which, in the worst case scenario gives

1. $\#Disc_{Sep.} = O(N)$
2. $\#Disc_{Clus.} = O(Nm)$, m number of atoms of the target
3. $\#Disc_{Intrp.} = O(Nm)$

Remark

If the input measures are discrete, with $n >> 1$ atoms, or n multiple of m, the estimates are independent of n!

Backpropagation / Adjoint method

Training: Backpropagation/Adjoint method

Assume that we have a differential equation

$$x(t)' = f(x(t), \theta(t))$$
$$x(0) = x_0$$

and we want to minimize

$$\min J(u) = \min \Phi(x(1)) + \frac{\varepsilon}{2} \int \theta(t)^2$$

Differentation

Differentiate the equation in the direction $\delta\theta$

$$\dot{x}'(t) = \partial_x f(x(t), \theta(t))\dot{x}(t) + \partial_\theta f(x(t), \theta(t))\delta\theta \tag{1}$$

Now let us differentiate the functional J

$$DJ(\theta)[\delta\theta] = \int_0^1 \langle \theta, \delta\theta \rangle \, \mathrm{d}t + \nabla_x \Phi(x(1))\dot{x}(1) \tag{2}$$

Introduce the adjoint equation (magic at first)

$$-p'(t) = \partial_x f(x(t), \theta(t))p(t)$$
$$p(t) = \nabla_x \Phi(x(1))$$

Manipulations

$$DJ(\theta)[\delta\theta] = \int_0^1 \langle \theta, \delta\theta \rangle \, \mathrm{d}t + \nabla_x \Phi(x(1))\dot{x}(1)$$

Plug the adjoint

$$DJ(\theta)[\delta\theta] = \int_0^1 \langle \theta, \delta\theta \rangle \, \mathrm{d}t + p(x(1))\dot{x}(1)$$

Then

$$DJ(u)[\delta\theta] = \int_0^1 \langle \theta, \delta\theta \rangle \, \mathrm{d}t + \int_0^1 (p(x(t))\dot{x}(t))' \, \mathrm{d}t + \underbrace{p(x(0))\dot{x}(0)}_{=0}$$

More manipulations

Then plugging the expressions we obtain

$$DJ(\theta)[\delta\theta] = \int_0^1 \langle \theta, \delta u \rangle \, \mathrm{d}t + \int_0^1 (p'(x(t))\dot{x}(t) + p(t)\dot{x}'(t)) \, \mathrm{d}t$$

Hence

$$DJ(\theta)[\delta\theta] = \int_0^1 \langle \theta + \partial_\theta f(x(t), \theta(t)), \delta\theta \rangle \, \mathrm{d}t$$

Therefore we can see that the gradient is equal to

$$\nabla J(\theta) = \theta + \partial_\theta f(x, \theta) p \tag{3}$$

Gradient descent

Set a learning rate ε
Solve

$$\frac{\mathrm{d}}{\mathrm{d}t}x^k(t) = f(x^k(t), \theta^k(t))$$
$$x^k(0) = x_0$$

and

$$-\frac{\mathrm{d}}{\mathrm{d}t}p^k(t) = \partial_x f(x^k(t), \theta^k(t))p(t)$$
$$p^k(1) = \nabla_x \Phi(x^k(1))$$

Update θ

$$\theta^{k+1} = \theta^k - \varepsilon\left(\theta^k + \partial_\theta f(x, \theta^k)p^k\right) \tag{4}$$

Matching disentangled measures

On the propagation of the assumption

Lemma

If

For any $1 \leq i \leq N$*, there exists* $\tilde{\mathrm{T}}^i \in L^2(\mathbb{S}^{d-1};\mathbb{S}^{d-1})$ *such that* $\tilde{\mathrm{T}}^i_{\#}\mu_0^i = \mu_1^i$.

Then

For any $1 \leq i \leq N$*, there exists* $\mathrm{T} \in L^2(\mathbb{S}^{d-1};\mathbb{S}^{d-1})$ *(common for all i!) such that* $\mathrm{T}_{\#}\Phi^T_{\theta_1}(\mu_0^i) = \Phi^T_{\theta_3}(\mu_1^i)$.

Inequality linking with Universal approximation

Lemma

Suppose $\mu \in \mathscr{P}(\mathbb{S}^{d-1})$ *and* $\mathrm{T}^1, \mathrm{T}^2 : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ *measurable, with* T^1 *bijective. Then*

$$W_2\left(\mathrm{T}^1_{\#}\mu, \mathrm{T}^2_{\#}\mu\right) \lesssim \left\|\mathrm{T}^1 - \mathrm{T}^2\right\|_{L^2(\mu)}.$$

Universal approximation

So, if we are able to approximate the common transport map T (in $L^2(\mu)$) with the flow of a Neural ODE (for instance) we are done!

Remark

When μ is AC, and T^1 and T^2 are the OTM between μ and ν_1, and μ and ν_2, The upper bound is know as the *linearized optimal transport distance*

[Delalande, Merigot, 2023]

Universal approximation

Lemma

Let $\varepsilon > 0$ *and* $\mu \in \mathscr{P}(\mathbb{S}^{d-1})$. *For every* $\mathsf{T} \in L^2(\mathbb{S}^{d-1}; \mathbb{S}^{d-1})$ *there exist a diffeomorphism* $\mathsf{T}_\varepsilon : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ *induced by the solution map of the Transformer (Neural ODE part), namely,*

$$\Phi^t_{\theta_\varepsilon}(\mu) = (\mathsf{T}_\varepsilon)_{\#}\mu$$

for some piecewise constant parameters θ_ε, *such that*

$$\|\mathsf{T} - \mathsf{T}_\varepsilon\|_{L^2(\mu)} \leq \varepsilon.$$

Universal approximation

The universal approximation will be based on

- Approximate first T by a piece-wise constant map Ψ^ε

$$\|\mathrm{T} - \Psi^\varepsilon\|_{L^2(\mu)} \leq \frac{\varepsilon}{2}$$

- Now we approximate Ψ^ε

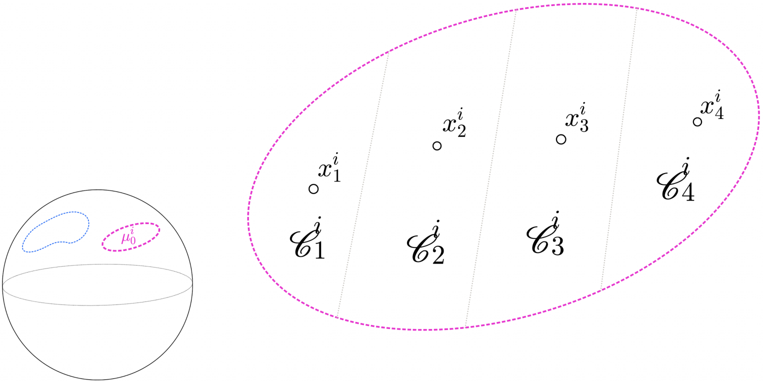
The key ingredients are

- Clustering
- Interpolation (simultaneous control) of M to M points in $\mathbb{S}^{d-1}$.

Clustering

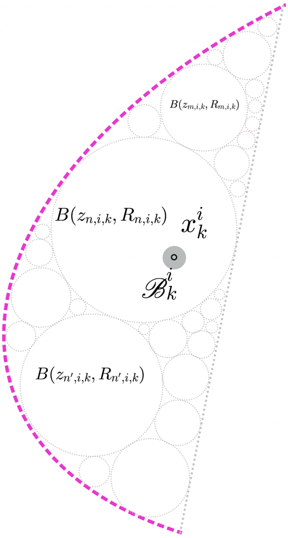
Clustering will allow to reduce the problem to an interpolation problem

Clustering



μ_0^i
x_1^i
x_2^i
x_3^i
x_4^i
$\mathscr{C}_1^i$
$\mathscr{C}_2^i$
$\mathscr{C}_3^i$
$\mathscr{C}_4^i$

Clustering



Lemma

Consider two open balls $\mathscr{B}_0, \mathscr{B}_1 \subset \mathbb{S}^{d-1}$ s.t. $\mathscr{B}_0 \cap \mathscr{B}_1 \neq \varnothing$. For any $\varepsilon > 0$ and $T > 0$, there exist $\boldsymbol{W}, \boldsymbol{V} \in \mathscr{M}_{d\times d}(\mathbb{R})$ and $b \in \mathbb{R}^d$ s.t. for any $\mu_0 \in \mathscr{P}(\mathbb{S}^{d-1})$, the solution $\mu \in \mathscr{C}^0([0,T]; \mathscr{P}(\mathbb{S}^{d-1}))$ satisfies

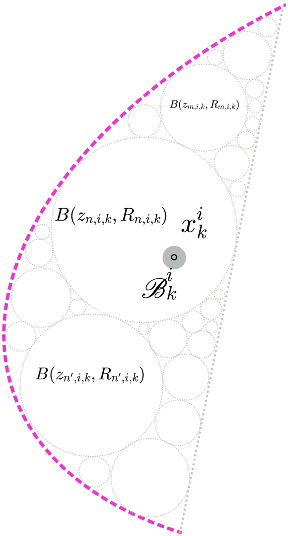
$$\mu(T, \mathscr{B}_0 \cap \mathscr{B}_1) \geq (1-\varepsilon)\mu_0(\mathscr{B}_0).$$

Moreover $\mu(T) = \Phi^T_{\#}\mu_0$ for a diffeomorphism $\Phi^t : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ which satisfies $(\Phi^t)_{|\mathscr{B}_0^c} \equiv \mathrm{Id}$ for $t \geq 0$.

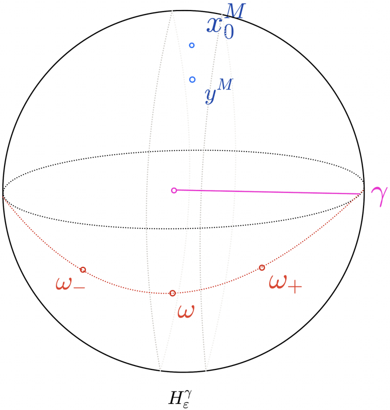
$$\dot{x}(t) = \mathbf{P}^{\perp}_{x(t)} \boldsymbol{W} \sigma(\boldsymbol{U} x(t) + b)$$

- $\boldsymbol{U}x + b$ is a hyperplane cutting the sphere
- $\boldsymbol{W}$ allows to choose a direction ω
- Thanks to the projection to the sphere, there is clustering to ω for the "activated" points.

Clustering



Interpolation



Interpolation

$$\Lambda = (\Psi_1)^{-1} \circ \Psi_2 \circ \Psi_1$$

satisfies

$$\Lambda(x_0^M) = y^M \qquad \Lambda(x_0^i) = x_0^i$$