



UNIVERSITAT DE
BARCELONA

Facultat de Matemàtiques
i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

Dinàmica de l'arquitectura Transformer: Del pas discret a la modelització contínua

Autor: Arshdeep Singh Kaur

Director: Dr. Domènec Ruiz-Balet

Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 10 de juny de 2026

Abstract

In this document we study the Transformer architecture from a mathematical viewpoint, interpreting it as a particle system, to model the trajectory of the tokens we derive a first order differential equation, and we construct an energy function to obtain a second order dynamics as the continuous limit of a method called Hamiltonian (Symplectic Euler). We relate this dynamics to Polyak's heavy-ball method and analyze its behavior. Finally, we validate the results with numerical simulations and explore the *clustering*.

Resum

En aquest treball estudiem matemàticament l'arquitectura *Transformer* que interpretem com un sistema de partícules, per modelar la trajectòria dels *tokens* derivem una equació diferencial de primer ordre i construïm una funció d'energia per obtenir una dinàmica de segon ordre a partir d'un esquema anomenat Hamiltonian (Euler simplèctic) i fer el seu límit continu. Relacionem aquesta dinàmica amb el mètode *heavy-ball* de Polyak i analitzem el comportament. Finalment, validem els resultats obtinguts amb simulacions numèriques i explorem el *clustering*.

Agraïments

Vull agrair l'ajuda i el suport durant els últims mesos al meu tutor Domènec Ruiz-Balet, i a la meva família i amics per ser-hi sempre.

Índex

1	Introducció i motivació	1
2	Com funciona un <i>Transformer</i>?	2
2.1	Del <i>transformer</i> a una equació diferencial	3
3	Cap a una dinàmica de segon ordre	6
3.1	Construcció de l'energia del sistema	6
3.2	L'esquema symplectic-Euler i el seu límit continu	11
4	Connexió amb <i>heavy-ball</i> i anàlisi de l'energia	15
5	Simulacions numèriques	24
6	Conclusions	41
A		42

Capítol 1

Introducció i motivació

La motivació d'aquest treball neix de la curiositat d'entendre el funcionament dels diferents models d'intel·ligència artificial que s'han popularitzat els últims anys. El primer contacte amb aquest tema va ser un seminari sobre matemàtiques a la intel·ligència artificial que vaig assistir poques setmanes abans de decidir el tema del treball amb el meu tutor, aquells dies es va despertar en mi un gran interès per aprendre sobre aquest ampli tema, i amb aquest objectiu, el meu tutor em va recomanar seguir la línia d'interpretar els *Transformers* com a sistemes de partícules de [2].

En aquest treball tenim com a objectiu estudiar matemàticament l'arquitectura *Transformer* interpretant-la com un sistema de partícules on cada *token* és una partícula, les capes passen a ser instants d'un temps continu, l'*Attention* i el MLP forces. Això ens permet modelar-ho com una equació diferencial i estudiar el comportament des d'un punt de vista dinàmic.

En relació a l'organització del document, al capítol 2 fem una introducció bàsica al funcionament dels *Transformers* i també derivem l'EDO de primer ordre que modela la trajectòria dels *tokens*. Al capítol 3 construïm la funció d'energia del sistema i obtenim una dinàmica de segon ordre a partir del límit continu de l'esquema de [4]. Al capítol 4 identifiquem aquesta dinàmica amb el mètode *heavy-ball* de Polyak i fem un anàlisi del comportament de l'energia i els punts d'equilibri. Per acabar, al capítol 5 validem els resultats obtinguts durant el treball amb simulacions numèriques i explorem el comportament que veiem en els diferents experiments.

Capítol 2

Com funciona un *Transformer*?

En aquest capítol tractarem de donar una introducció bàsica de com funcionen els *Transformers*, l'arquitectura introduïda per [1] que hi ha al darrere dels diferents *LLM* actuals, com poden ser GPT-5, Gemini o Claude.

Primer de tot necessitem entendre que és un *token*, els *transformers* comencen processant text (normalment) i el divideixen en paraules o fragments més petits, que anomenem *tokens*. Aquests estan representats per un *embedding*, és a dir, un vector de $\mathbb{R}^d$, en cas de tenir n tokens es representen en una matriu $X \in \mathbb{R}^{n \times d}$ on cada fila és un *embedding*. El que fa el *transformer* és passar aquesta matriu per diverses capes de la xarxa amb l'objectiu de convertir-la en una representació final que sigui útil per la tasca que es vol fer, per exemple predir la següent paraula d'una frase. Les dues operacions que es fan quan passen per les capes són l'*Attention* i el MLP.

L'*Attention* és la tècnica que relaciona cada token amb els altres per actualitzar-se, per fer-ho utilitza tres vectors relacionats a cada token. *Query* (Q) representa la informació que el token està buscant en els altres, *Key* (K) la informació que un token pot oferir i *Value* (V) la informació que es transmet cap a cada token finalment. Aquest últim s'obté comparant les *queries* amb les *keys*, esbrinant així quanta atenció ha de prestar cada token als altres. Per fer això s'utilitza el *softmax*, que transforma les puntuacions en una distribució de probabilitat sobre els n tokens. Per a cada token i definim la funció de partició

$$Z_{\beta,i} = \sum_{k=1}^n e^{\beta \langle Qx_i, Kx_k \rangle}$$

i els pesos d'*Attention*

$$A_{ij} = \frac{e^{\beta \langle Qx_i, Kx_j \rangle}}{Z_{\beta,i}}$$

de manera que A_{ij} calcula l'atenció del token i al token j . L'*Attention* de cada token és aleshores

$$\text{Attn}(X)_i = \sum_{j=1}^n A_{ij} Vx_j$$

on Q, K, V són matrius apreses durant l'entrenament i $\beta > 0$ fa el paper de $1/\sqrt{d}$ del *transformer* original, és el mateix *softmax attention*.

Una vegada el token ha passat per la fase de l'*Attention*, apareix el MLP (*Multilayer Perceptron*), que només depèn del token individualment (a diferència de l'*Attention*). Habitualment, consisteix en dues capes lineals amb una funció d'activació. La capa es defineix com una aplicació $z \mapsto Wz + b$ on W és una matriu de pesos i b un vector. La funció d'activació és una funció no lineal $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ que s'aplica a cada component d'un vector, $\sigma(v) = (\sigma(v_1), \dots, \sigma(v_m))$. S'utilitza per donar al MLP caràcter no lineal. Un exemple popular és ReLU, $\sigma(z) = \max(0, z)$. Llavors el MLP és

$$\text{MLP}(x) = W_2 \sigma(W_1 x + b_1) + b_2$$

Per aprofundir més matemàticament en els MLP es pot consultar [12].

Els transformers consisteixen en un conjunt de blocs, on primer sumem l'*Attention* al token (i passen per una normalització *layer normalization*, que ometrem per preservar l'estructura que estudiarem) i al resultat obtingut li sumem el MLP. Això es pot representar com $X_1 = X_0 + \text{Attn}(X_0)$ i $X_2 = X_1 + \text{MLP}(X_1)$. Cada bloc pot tenir els seus propis pesos (apresos durant l'entrenament), la matriu inicial X va passant per aquests fins arribar al seu estat final.

Ara que ja hem explicat com funciona un *Transformer*, centrem-nos en la versió que motiva aquest treball, que consisteix en veure'l com a sistema de partícules. Aquesta idea està introduïda a [2]. Veiem cada *token* com a partícula, l'*Attention* com una força d'interacció i el MLP com una força que actua sobre cada partícula per separat. El temps fa el rol de l'índex de les capes i així es pot interpretar tot com que les partícules es mouen per $\mathbb{R}^d$ amb aquestes forces. Aquest moviment a l'espai està regulat per una EDO, ja que en lloc de moure's capa a capa, es mouen de forma contínua en el temps.

2.1 Del *transformer* a una equació diferencial

En aquesta secció ens centrem en com evoluciona la trajectòria d'un token al llarg de les diferents capes. Ho fem per un sol token perquè l'estructura es preserva en el cas general de n tokens, que explorarem més endavant.

Partint del pseudocodi el bloc es redueix a

$$X_1 = X_0 + \text{Attn}(X_0), \quad X_2 = X_1 + \text{MLP}(X_1) \quad (2.1.1)$$

passant per N blocs d'aquest tipus obtenim tota la xarxa.

En lloc d'entendre les capes com passos discrets, les interpretem com instants d'un temps continu t com hem mencionat en la secció anterior. Seguint l'enfocament de [3, A. Deriving the model], prescindint de les capes de normalització (*LayerNorm*) que apareixen a la implementació real obtindrem l'EDO que modelitza la trajectòria per tota la xarxa.

Els blocs del *transformer* estan numerats per $t \in \{1, 2, \dots, N\}$. Per veure-ho com un límit continu, hem de fer que t prengui valors en un interval $[0, \infty)$ i que

les capes siguin instants d'aquesta evolució. Suposem que existeix una trajectòria suau $X : [0, \infty) \rightarrow \mathbb{R}^d$ tal que la t -èsima capa correspon al valor $X(t)$. Passar d'una capa a la següent equival a avançar un temps $\Delta t = 1$, que nosaltres farem anar a zero.

Volem ara reescriure el pseudocodi amb el paràmetre Δt tal que per a $\Delta t = 1$ sigui equivalent a (2.1.1), i a la vegada, fent $\Delta t \rightarrow 0$ obtinguem una EDO contínua.

Per fer-ho d'una manera coherent, multipliquem Δt per les contribucions de l'*Attention* i el MLP respectivament a X_1 i X_2 . Amb això obtenim

$$X_1 = X_0 + \Delta t \text{Attn}(X_0), \quad X_2 = X_1 + \Delta t \text{MLP}(X_1) \quad (2.1.2)$$

Substituint X_1 a X_2 ,

$$X_2 = X_0 + \Delta t \text{Attn}(X_0) + \Delta t \text{MLP}(X_0 + \Delta t \text{Attn}(X_0)) \quad (2.1.3)$$

Reanomenant $X_0 \mapsto X_t$ i $X_2 \mapsto X_{t+1}$ obtenim l'expressió del bloc sense passos intermitjos.

$$X_{t+1} = X_t + \Delta t \text{Attn}(X_t) + \Delta t \text{MLP}(X_t + \Delta t \text{Attn}(X_t)) \quad (2.1.4)$$

És senzill veure que si $\Delta t = 1$ això últim és equivalent al bloc original, ja que els tres Δt desapareixen, sent només una reescriptura i no un model diferent.

Parem un moment per recordar que per manipular $\mathcal{O}(\Delta t^k)$ quan fem $\Delta t \rightarrow 0$ seguim

$$\mathcal{O}(\Delta t^a) + \mathcal{O}(\Delta t^b) = \mathcal{O}(\Delta t^{\min(a,b)}) \quad \text{i} \quad \Delta t^a \cdot \mathcal{O}(\Delta t^b) = \mathcal{O}(\Delta t^{a+b})$$

per sumar i multiplicar, amb $a, b \in \mathbb{N}$.

Ens queda comprovar que per a $\Delta t \rightarrow 0$ obtenim una EDO contínua.

A l'expressió (2.1.4) veiem que MLP s'avalua a $X_t + \Delta t \text{Attn}(X_t)$. Per poder veure com es comporta al límit $\Delta t \rightarrow 0$, ho expandim per Taylor al voltant de X_t .

$$\text{MLP}(X_t + \Delta t \text{Attn}(X_t)) = \text{MLP}(X_t) + \Delta t \nabla \text{MLP}(X_t) \text{Attn}(X_t) + \mathcal{O}(\Delta t^2) \quad (2.1.5)$$

sempre que MLP sigui $\mathcal{C}^2$, que assumirem al llarg de tot el treball. Substituint (2.1.5) dins (2.1.4) obtenim

$$X_{t+1} = X_t + \Delta t \text{Attn}(X_t) + \Delta t \text{MLP}(X_t) + \Delta t^2 \nabla \text{MLP}(X_t) \text{Attn}(X_t) + \mathcal{O}(\Delta t^3)$$

i agrupant els termes del mateix ordre,

$$X_{t+1} = X_t + \Delta t (\text{Attn}(X_t) + \text{MLP}(X_t)) + \mathcal{O}(\Delta t^2) \quad (2.1.6)$$

Com després farem $\Delta t \rightarrow 0$ per trobar l'equació diferencial, els termes d'ordre Δt^2 són negligibles. Restant X_t a banda i banda de (2.1.6) i dividint per Δt obtenim

$$\frac{X^{t+1} - X_t}{\Delta t} = \text{Attn}(X_t) + \text{MLP}(X_t) + \mathcal{O}(\Delta t)$$

Fent el límit $\Delta t \rightarrow 0$ el membre esquerre és per definició la derivada de $X(t)$ i el terme d'error desapareix. Obtenim així el que buscàvem, que coincideix amb [4].

Teorema 2.1.1. *La trajectòria $X(t) \in \mathbb{R}^d$ d'un token satisfà*

$$\frac{dX}{dt} = \text{Attn}(X(t)) + \text{MLP}(X(t)) \quad (2.1.7)$$

Notem que no hi ha res que controli el creixement de $\|X(t)\|$, que podria divergir depenent de com es comportin Attn i MLP. També cal remarcar que només hem treballat amb un sol token $X \in \mathbb{R}^d$ per simplicitat i seguir fàcilment la trajectòria d'un token, la demostració s'estén al cas $X \in \mathbb{R}^{n \times d}$ amb $n \geq 1$ tokens, ja que l'únic que hem fet servir són la suma i el producte per escalars, que $\text{MLP} \in \mathcal{C}^2$ (per al desenvolupament de Taylor) i el pas al límit $\Delta t \rightarrow 0$, cap de les quals depèn de la dimensió. A partir d'ara treballarem en aquest cas i estudiarem un model de segon ordre que controli aquest creixement.

Capítol 3

Cap a una dinàmica de segon ordre

3.1 Construcció de l'energia del sistema

Per resoldre el problema de controlar el creixement que hem vist al capítol anterior, necessitem introduir el *moment* V .

Definició 3.1.1. *Anomenem moment del conjunt de tokens a l'instant t a $V(t) \in \mathbb{R}^{n \times d}$ definit per*

$$V(t) := \frac{dX(t)}{dt}$$

Notem que habitualment el moment és $p = mV$, on m és la massa i V la velocitat. Aquí tots els *tokens* són objectes idèntics, jugant el mateix paper en la dinàmica i per tant els hi assignem la mateixa "massa". Per simplicitat triem $m = 1$. Llavors el moment i la velocitat coincideixen, $V = \dot{X}$.

Volem construir una funció $H : \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times d} \rightarrow \mathbb{R}$ que representi l'energia total del sistema. A la mecànica clàssica aquesta energia es descompon en energia cinètica K , dependent del moment V i potencial U , dependent de la posició X .

$$H(X, V) = K(V) + U(X) \quad (3.1.1)$$

Comencem escollint l'energia cinètica habitual $K(V) = \frac{1}{2} \|V\|_F^2$, amb la norma de Frobenius $\|V\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^d |v_{ij}|^2}$, la utilitzarem per totes les matrius durant el treball. Per a l'energia potencial, com volem que l'Attention i el MLP actuïn com a forces del sistema, hem d'escollir un $U(X) \in \mathcal{C}^2$ tal que

$$\nabla U(X) = \text{Attn}(X) + \text{MLP}(X) \quad (3.1.2)$$

Anem a veure que existeix una funció $U : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}$ de classe $\mathcal{C}^2$ tal que

$$\nabla U(X) = \text{Attn}(X) + \text{MLP}(X) \quad \text{per a tot } X \in \mathbb{R}^{n \times d} \quad (3.1.3)$$

Notem que pel teorema de Schwarz, si $\text{Attn} + \text{MLP} = \nabla U$ amb $U \in \mathcal{C}^2$, llavors la matriu jacobiana (és la hessiana de U) de $\text{Attn} + \text{MLP}$ ha de ser simètrica a tot punt. Pel *Attention* notem que aquesta jacobiana és simètrica quan $Q^\top K$ és simètrica definida positiva (amb $V = Q^\top K$) com es veu a [2], com utilitzem $Q = K = V = I_d$ això es compleix.

Buscarem U de la forma (3.1.2) com

$$U(X) = U_{\text{Attn}} + U_{\text{MLP}} \quad (3.1.4)$$

Veiem que existeix un U com el que hem comentat, prenem com a *Attention* la *unnormalized self attention* de [2] amb el paràmetre β i $Q = K = V = I$. Per tant,

$$\text{Attn}(X)_i = \frac{1}{n} \sum_{j=1}^n e^{\beta \langle x_i, x_j \rangle} x_j \quad (3.1.5)$$

Utilitzem la versió sense normalitzar per $\sum_j e^{\beta \langle x_i, x_j \rangle}$ perquè si ho fèssim no és compliria la condició de ser gradient. L'*Attention* (3.1.5) prové de l'energia d'interacció de [2]

$$E_\beta(X) = \frac{1}{2\beta n^2} \sum_{i,j=1}^n e^{\beta \langle x_i, x_j \rangle}$$

Multiplicant per n ,

$$U_{\text{Attn}}(X) = n E_\beta(X) = \frac{1}{2\beta n} \sum_{i,j=1}^n e^{\beta \langle x_i, x_j \rangle} \quad (3.1.6)$$

per tal que el gradient doni l'atenció, $\nabla U_{\text{Attn}} = \text{Attn}$ (el gradient de E_β tal qual seria $\frac{1}{n} \text{Attn}$, hem fet això per adaptar el que es diu al paper a les nostres condicions).

Ens centrem a comprovar que $\nabla U_{\text{Attn}} = \text{Attn}(X)$, comencem derivant el producte escalar

$$\frac{\partial \langle x_i, x_j \rangle}{\partial x_k} = \delta_{ik} x_j + \delta_{jk} x_i$$

seguim amb

$$\begin{aligned} \nabla_{x_k} U_{\text{Attn}} &= \frac{1}{2\beta n} \sum_{i,j} e^{\beta \langle x_i, x_j \rangle} \beta (\delta_{ik} x_j + \delta_{jk} x_i) = \frac{1}{2n} \left(\sum_j e^{\beta \langle x_k, x_j \rangle} x_j + \sum_i e^{\beta \langle x_i, x_k \rangle} x_i \right) \\ &= \frac{1}{n} \sum_j e^{\beta \langle x_k, x_j \rangle} x_j = \text{Attn}(X)_k \end{aligned}$$

on hem fet servir que els dos sumatoris són iguals per simetria de $\langle x_i, x_j \rangle = \langle x_j, x_i \rangle$.

Ara intentem buscar un U_{MLP} , si aquest ve d'un MLP que actua component a component amb pesos diferents com

$$\text{MLP}(X)_{ij} = -\alpha_i X_{ij}, \quad U_{\text{MLP}}(X) = -\sum_{i=1}^n \frac{\alpha_i}{2} \|x_i\|^2 \quad (3.1.7)$$

Comprovem $\nabla U_{\text{MLP}}(X) = \text{MLP}(X)$ derivant cada component $\frac{\partial U_{\text{MLP}}}{\partial X_{kl}} = -\alpha_k X_{kl}$, això garanteix les suposicions es continuen complint ja que tenim $U_{\text{total}} = U_{\text{Attn}} + U_{\text{MLP}}$ i el gradient d'una suma és la suma dels gradients.

Aquest exemple de MLP és lineal, un cas senzill per veure que es compleixen les condicions, en general el bloc del MLP acostuma a tenir una part no lineal σ (com ReLU o GeLU) llavors,

$$\text{MLP}(X)_i = \sum_j \omega_j \sigma(a_j \cdot x_i) \quad a_j, \omega_j \in \mathbb{R}^d$$

A [10] es veu que aquest MLP deriva d'un potencial si $\omega_j = c_j a_j$ amb $c_j \in \mathbb{R}$. Si tenim una primitiva φ tal que $\varphi' = 2\sigma$, el potencial és

$$U_{\text{MLP}}(X) = \frac{1}{2} \sum_i \sum_j c_j \varphi(a_j \cdot x_i)$$

ja que $\nabla_{x_i} U_{\text{MLP}} = \frac{1}{2} \sum_j c_j \varphi'(a_j \cdot x_i) a_j = \sum_j c_j a_j \sigma(a_j \cdot x_i) = \text{MLP}(X)_i$

Amb això veiem que tant l'exemple de MLP lineal que hem vist com aquest deriven d'un potencial, que és el que volíem.

Ara posarem un altre $\text{Attn}(X)$ d'exemple que compleix les condicions que hem dit abans. Per β petita expandim $e^{\beta \langle x_i, x_j \rangle} = 1 + \beta \langle x_i, x_j \rangle + O(\beta^2)$ de (3.1.5) i obtenim

$$\text{Attn}(X)_i = \frac{1}{n} \sum_{j=1}^n e^{\beta \langle x_i, x_j \rangle} x_j = \frac{1}{n} \sum_{j=1}^n x_j + \beta \frac{1}{n} \sum_{j=1}^n \langle x_i, x_j \rangle x_j + O(\beta^2)$$

On el primer terme $\frac{1}{n} \sum_{j=1}^n x_j$ és el centroide i no depèn de x_i , el segon coeficient de β sí que depèn de x_i i és el que utilitzarem com a *Attention*. Notem que si prenem el *linear Attention* de [9] amb $Q = K = V = I_d$ (com hem fet fins ara), no utilitzem el *feature map* tenim

$$\text{Attn}_{\text{linear}}(x_i) = \frac{\phi(Qx_i)^\top \sum_{j=1}^n \phi(Kx_j) Vx_j^\top}{\phi(Qx_i)^\top \sum_{j=1}^n \phi(Kx_j)} = \frac{\phi(x_i)^\top \sum_{j=1}^n \phi(x_j) x_j^\top}{\phi(x_i)^\top \sum_{j=1}^n \phi(x_j)} = \frac{x_i^\top \sum_{j=1}^n x_j x_j^\top}{x_i^\top \sum_{j=1}^n x_j} =$$

passem $x_i^\top$ dintre del sumatoris i fem servir que $x_i^\top x_j = \langle x_i, x_j \rangle$

$$= \frac{\sum_{j=1}^n x_i^\top x_j x_j^\top}{\sum_{j=1}^n x_i^\top x_j} = \frac{\sum_{j=1}^n \langle x_i, x_j \rangle x_j^\top}{\sum_{j=1}^n \langle x_i, x_j \rangle}$$

I substituint el denominador per n per que pugui tenir potencial, arribem al mateix *Attention*.

$$\text{Attn}(x_i) = \frac{1}{n} \sum_{j=1}^n \langle x_i, x_j \rangle x_j \quad (3.1.8)$$

Hem fet això per veure que aquesta versió està fortament relacionada amb l'anterior (3.1.5) (assignant β a prop de 0) i amb la linear Attention. Cal destacar que la part del nom *linear* ve de la seva complexitat computacional i no perquè sigui lineal en el sentit matemàtic.

Ara podem expressar (3.1.8) en forma matricial prenent $X \in \mathbb{R}^{n \times d}$ i els *tokens* com a files

$$\text{Attn}(X) = \frac{1}{n} X X^T X \quad (3.1.9)$$

D'aquesta manera només tenint en compte l'Attention, el potencial és

$$U_{\text{Attention}}(X) = \frac{1}{4n} \sum_{i,j} \langle x_i, x_j \rangle^2 = \frac{1}{4n} \|X X^T\|_F^2 \quad (3.1.10)$$

i $\nabla_X U(X) = \text{Attn}(X)$. Demostrem-ho.

Demostració. Definim $A := X X^T \in \mathbb{R}^{n \times n}$ aquesta matriu és simètrica, ja que $A^T = (X X^T)^T = X X^T = A$ i per la norma de Frobenius,

$$\|X X^T\|_F^2 = \|A\|_F^2 = \sum_{i,j} A_{ij}^2 = \sum_{i,j} A_{ij} A_{ji} = \text{tr}(A^2)$$

a la penúltima igualtat hem utilitzat la simetria $A_{ji} = A_{ij}$ i a l'última que $\text{tr}(A^2) = \sum_i (A^2)_{ii} = \sum_{i,j} A_{ij} A_{ji}$. Per tant

$$U(X) = \frac{1}{4n} \text{tr}(A^2)$$

Ara calculem $\nabla_X U(X)$ component a component, sigui $k \in \{1, \dots, n\}$ i $l \in \{1, \dots, d\}$. Volem veure

$$\frac{\partial U}{\partial X_{kl}} = \frac{1}{4n} \frac{\partial \text{tr}(A^2)}{\partial X_{kl}}$$

Reescrivint $\text{tr}(A^2) = \sum_{i,j} A_{ij}^2$ i aplicant la regla de la cadena tenim

$$\frac{\partial \text{tr}(A^2)}{\partial X_{kl}} = \sum_{i,j} 2A_{ij} \frac{\partial A_{ij}}{\partial X_{kl}} \quad (3.1.11)$$

Necessitem calcular la derivada parcial de cada entrada A_{ij} respecte a X_{kl} . Utilitzant que

$$A_{ij} = (X X^T)_{ij} = \sum_{p=1}^d X_{ip} X_{jp}$$

i aplicant la regla del producte tenint en compte que $\frac{\partial X_{ip}}{\partial X_{kl}} = \delta_{ik} \delta_{pl}$

$$\frac{\partial A_{ij}}{\partial X_{kl}} = \sum_p (\delta_{ik} \delta_{pl} X_{jp} + X_{ip} \delta_{jk} \delta_{pl}) = \delta_{ik} X_{jl} + \delta_{jk} X_{il} \quad (3.1.12)$$

on δ_{ab} és la delta de Kronecker ($\delta_{ab} = 1$ si $a = b$, altrament és 0).

Substituint (3.1.12) a (3.1.11),

$$\frac{\partial \text{tr}(A^2)}{\partial X_{kl}} = 2 \sum_{i,j} A_{ij} \delta_{ik} X_{jl} + 2 \sum_{i,j} A_{ij} \delta_{jk} X_{il}$$

Als dos sumatoris veiem que la delta de Kronecker redueix i o j a només un valor. Fixem-nos en que

- Al primer sumatori δ_{ik} implica $i = k$ i la suma sobre i es redueix a un sol terme amb $i = k$

$$2 \sum_j A_{kj} X_{jl}$$

- Al segon δ_{jk} implica $j = k$ i la suma sobre j es redueix a un sol terme amb $j = k$ i fent servir que $A_{ik} = A_{ki}$.

$$2 \sum_i A_{ik} X_{il} = 2 \sum_i A_{ki} X_{il}$$

Si reanomenem l'índex i a j

$$\frac{\partial \text{tr}(A^2)}{\partial X_{kl}} = 2 \sum_j A_{kj} X_{jl} + 2 \sum_j A_{kj} X_{jl} = 4 \sum_j A_{kj} X_{jl}$$

Pel producte de matrius $\sum_j A_{kj} X_{jl} = (AX)_{kl}$ i llavors

$$\frac{\partial \text{tr}(A^2)}{\partial X_{kl}} = 4(AX)_{kl} \quad (3.1.13)$$

Veiem que això últim és equivalent a

$$\frac{\partial \text{tr}((XX^T)^2)}{\partial X} = 4AX = 4XX^T X$$

Per tant tenim

$$\nabla_X U(X) = \frac{1}{4n} \frac{\partial \text{tr}((XX^T)^2)}{\partial X} = \frac{1}{4n} 4XX^T X = \frac{1}{n} XX^T X = \text{Attn}(X)$$

com volíem demostrar. $\square$

Amb tot això queda clar que existeix la funció d'energia $H(X, V) = \frac{1}{2} \|V\|_F^2 + U(X)$ tal que $\nabla U(X) = \text{Attn}(X) + \text{MLP}(X)$.

A la secció següent calcularem el límit continu del mètode *symplectic Euler* de [4].

3.2 L'esquema symplectic-Euler i el seu límit continu

Acabem de proposar una funció $H(X, V)$ que representa l'energia del sistema i hem demostrat l'existència del potencial $U \in \mathcal{C}^2$ tal que $\nabla U(X) = \text{Attn}(X) + \text{MLP}$. Seguirem avançant per trobar una dinàmica de segon ordre explorant el Hamiltonian Symplectic-Euler de [4]. Ens centrarem en descriure'l, reparametritzar coeficients en funció de Δt , i calcular el sistema d'EDOs quan fem $\Delta t \rightarrow 0$.

Primer de tot introduïm que és un esquema simplèctic. Si tenim un sistema Hamiltonià partit

$$\dot{X} = V \quad \text{i} \quad \dot{V} = -\nabla U(X)$$

el *Symplectic-Euler* clàssic de [8] discretitza el sistema utilitzant un pas Δt , actualitzant primer la velocitat i després la posició de la següent manera,

$$V_{k+1} = V_k - \Delta t \nabla U(X_k), \quad X_{k+1} = X_k + \Delta t V_{k+1}$$

Un pas de l'esquema és una aplicació $\Phi_{\Delta t} : (X_k, V_k) \mapsto (X_{k+1}, V_{k+1})$, l'esquema és simplèctic quan en aquesta aplicació es preserva l'estructura geomètrica de l'espai de fases de manera que si prenem com M la seva jacobiana, s'ha de complir que

$$M^\top J M = J \quad \text{i} \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

Aquest mètode *Hamiltonian Symplectic-Euler* està basat en el que hem explicat per sistemes hamiltonians partits. El fem servir perquè encaixa amb el potencial U que hem buscat $-\nabla U = -(\text{Attn} + \text{MLP})$ i l'esquema afegix una variable de moment V per obtenir una dinàmica de segon ordre. A més, com afegim la fricció, el sistema passa a dissipar energia que és el que necessitem per controlar el creixement. Aquesta variant dissipativa fa que el mètode no sigui simplèctic.

Aquest mètode evoluciona durant tres passos seguits, aquests consisteixen en dos *kicks* i un *drift* entre aquests.

- Un *kick* actualitza el moment V aplicant una “força” (*Attn* o *MLP*) avaluada a $X + \mu V$ (un *lookahead* de mida μ).
- Un *drift* actualitza la posició X , avançant-la en la direcció del moment.

L'esquema que seguirem de [4] és:

$$\begin{aligned} \textbf{Attn-kick:} \quad & X_t^{\text{in}} = X_t + \mu_t V_t \\ & V_{t+\frac{1}{2}} = \beta_t V_t - \gamma_t \text{Attn}_t(X_t^{\text{in}}) \\ \textbf{Drift:} \quad & X_{t+\frac{1}{2}} = X_t + \delta_t V_{t+\frac{1}{2}} \\ \textbf{MLP-kick:} \quad & X_{t+\frac{1}{2}}^{\text{in}} = X_{t+\frac{1}{2}} + \mu_t V_{t+\frac{1}{2}} \\ & V_{t+1} = \beta_t V_{t+\frac{1}{2}} - \gamma_{t+\frac{1}{2}} \text{MLP}_t(X_{t+\frac{1}{2}}^{\text{in}}) \\ & X_{t+1} = X_{t+\frac{1}{2}} \end{aligned} \tag{3.2.1}$$

Els paràmetres $\mu_t, \delta_t, \gamma_t \in \mathbb{R}$ controlen els sub-passos i $\beta_t \in \mathbb{R}$ és el coeficient de fricció. Ara necessitem assignar com escalen per poder donar-li sentit al límit $\Delta t \rightarrow 0$.

$$\delta_t = \Delta t, \quad \mu_t = \mu' \Delta t, \quad \gamma_t = \gamma' \Delta t, \quad \beta_t = 1 - \beta' \Delta t, \quad \gamma_{t+\frac{1}{2}} = \gamma'' \Delta t \quad (3.2.2)$$

on tenim que $\mu', \gamma', \beta', \gamma'' \in \mathbb{R}$ són constants independents de Δt majors a 0 i $\beta' < 1$.

Tots aquests paràmetres porten Δt , el *Drift* avança la posició X en el temps segons el moment i δ_t , els *Lookahead* (línies amb X_t^{in}) semblen tenir el mateix propòsit però només s'encarreguen de fer un desplaçament auxiliar petit $\mu' \Delta t$. Els *Kicks* varien el moment segons γ', γ'' en el cas de l'*Attention* i el MLP respectivament. També resaltem l'existència de la fricció $\beta_t = 1 - \beta' \Delta t$, l'hem escollit així per imposar que si $\Delta t = 0$ no hi hagi dissipament, ja que $\beta_t = 1$.

Amb la reparametrització (3.2.2) tenim que (3.2.1) es converteix en:

$$\begin{aligned} \textbf{Attn-kick:} \quad & X_t^{\text{in}} = X_t + \mu' \Delta t V_t \\ & V_{t+\frac{1}{2}} = (1 - \beta' \Delta t) V_t - \gamma' \Delta t \text{Attn}(X_t^{\text{in}}) \\ \textbf{Drift:} \quad & X_{t+\frac{1}{2}} = X_t + \Delta t V_{t+\frac{1}{2}} \\ \textbf{MLP-kick:} \quad & X_{t+\frac{1}{2}}^{\text{in}} = X_{t+\frac{1}{2}} + \mu' \Delta t V_{t+\frac{1}{2}} \\ & V_{t+1} = (1 - \beta' \Delta t) V_{t+\frac{1}{2}} - \gamma'' \Delta t \text{MLP}(X_{t+\frac{1}{2}}^{\text{in}}) \\ & X_{t+1} = X_{t+\frac{1}{2}} \end{aligned} \quad (3.2.3)$$

Per obtenir el límit continu farem Taylor sobre cada línia de (3.2.3).

Suposem que *Attn* i *MLP* són camps vectorials $\mathcal{C}^2$ sobre $\mathbb{R}^{n \times d}$ per tal que els desenvolupaments de Taylor siguin vàlids.

Comencem amb el *Lookahead* d'*Attn-kick*,

$$X_t^{\text{in}} = X_t + \mu' \Delta t V_t \implies X_t^{\text{in}} - X_t = \underbrace{\mu' \Delta t V_t}_{\mathcal{O}(\Delta t)} \quad (3.2.4)$$

Expandint *Attn* per Taylor al voltant de X_t tenim

$$\text{Attn}(X_t^{\text{in}}) = \text{Attn}(X_t) + \mu' \Delta t V_t \nabla \text{Attn}(X_t) + \mathcal{O}(\Delta t^2) = \text{Attn}(X_t) + \mathcal{O}(\Delta t) \quad (3.2.5)$$

Ara si substituïm (3.2.5) a la segona línia de (3.2.3)

$$\begin{aligned} V_{t+\frac{1}{2}} &= (1 - \beta' \Delta t) V_t - \gamma' \Delta t (\text{Attn}(X_t) + \mu' \Delta t V_t \nabla \text{Attn}(X_t) + \mathcal{O}(\Delta t^2)) \\ &= (1 - \beta' \Delta t) V_t - \gamma' \Delta t \text{Attn}(X_t) + \mathcal{O}(\Delta t^2) \end{aligned} \quad (3.2.6)$$

Desenvolupant $(1 - \beta' \Delta t) V_t = V_t - \beta' \Delta t V_t$

$$V_{t+\frac{1}{2}} = V_t + \mathcal{O}(\Delta t) \quad (3.2.7)$$

Més endavant necessitarem l'expressió (3.2.6).

Pel *Drift* combinem la tercera línia de (3.2.3) amb (3.2.6)

$$\begin{aligned} X_{t+\frac{1}{2}} &= X_t + \Delta t V_{t+\frac{1}{2}} \\ &= X_t + \Delta t [(1 - \beta' \Delta t) V_t - \gamma' \Delta t \text{Attn}(X_t) + \mathcal{O}(\Delta t^2)] \\ &= X_t + \Delta t V_t + \mathcal{O}(\Delta t^2) \end{aligned} \quad (3.2.8)$$

Ara utilitzem la primera línia del bloc *MLP-kick* de (3.2.3) combinada amb (3.2.8) i (3.2.7) pel segon lookahead, tenim:

$$\begin{aligned} X_{t+\frac{1}{2}}^{\text{in}} &= X_{t+\frac{1}{2}} + \mu' \Delta t V_{t+\frac{1}{2}} \\ &= [X_t + \Delta t V_t + \mathcal{O}(\Delta t^2)] + \mu' \Delta t [V_t + \mathcal{O}(\Delta t)] \\ &= X_t + (1 + \mu') \Delta t V_t + \mathcal{O}(\Delta t^2) \end{aligned} \quad (3.2.9)$$

Expandint MLP per Taylor al voltant de X_t

$$\text{MLP}(X_{t+\frac{1}{2}}^{\text{in}}) = \text{MLP}(X_t) + (1 + \mu') \Delta t V_t \nabla \text{MLP}(X_t) + \mathcal{O}(\Delta t^2) = \text{MLP}(X_t) + \mathcal{O}(\Delta t) \quad (3.2.10)$$

Substituint (3.2.6) i (3.2.10) a la segona línia del bloc *MLP-kick* de (3.2.3) tenim:

$$\begin{aligned} V_{t+1} &= (1 - \beta' \Delta t) V_{t+\frac{1}{2}} - \gamma'' \Delta t \text{MLP}(X_{t+\frac{1}{2}}^{\text{in}}) \\ &= (1 - \beta' \Delta t) [(1 - \beta' \Delta t) V_t - \gamma' \Delta t \text{Attn}(X_t) + \mathcal{O}(\Delta t^2)] \\ &\quad - \gamma'' \Delta t [\text{MLP}(X_t) + \mathcal{O}(\Delta t)] \end{aligned}$$

Desenvolupem,

$$(1 - \beta' \Delta t)^2 = 1 - 2\beta' \Delta t + (\beta' \Delta t)^2 = 1 - 2\beta' \Delta t + \mathcal{O}(\Delta t^2)$$

i

$$(1 - \beta' \Delta t)(-\gamma' \Delta t) = -\gamma' \Delta t + \mathcal{O}(\Delta t^2)$$

$$V_{t+1} = (1 - 2\beta' \Delta t) V_t - \gamma' \Delta t \text{Attn}(X_t) - \gamma'' \Delta t \text{MLP}(X_t) + \mathcal{O}(\Delta t^2) \quad (3.2.11)$$

Agrupem,

$$V_{t+1} - V_t = -2\beta' \Delta t V_t - \Delta t (\gamma' \text{Attn}(X_t) + \gamma'' \text{MLP}(X_t)) + \mathcal{O}(\Delta t^2) \quad (3.2.12)$$

La última línia de (3.2.3) i (3.2.8) donen

$$X_{t+1} = X_{t+\frac{1}{2}} = X_t + \Delta t V_t + \mathcal{O}(\Delta t^2) \quad (3.2.13)$$

Per acabar dividim (3.2.12) i (3.2.13) per Δt .

$$\begin{aligned} \frac{X_{t+1} - X_t}{\Delta t} &= V_t + \mathcal{O}(\Delta t) \\ \frac{V_{t+1} - V_t}{\Delta t} &= -2\beta' V_t - (\gamma' \text{Attn}(X_t) + \gamma'' \text{MLP}(X_t)) + \mathcal{O}(\Delta t^2) \end{aligned}$$

En el límit $\Delta t \rightarrow 0$, els termes d'error s'anul·len i els membres esquerres són per definició les derivades respecte a t de X i V .

Teorema 3.2.1. *Sota la reparametrització (3.2.2) de l'esquema (3.2.3) (X_t, V_t) convergeixen a la solució del sistema d'EDOs per $\Delta t \rightarrow 0$*

$$\begin{cases} \dot{X} = V, \\ \dot{V} = -2\beta' V - (\gamma' \text{Attn}(X) + \gamma'' \text{MLP}(X)) \end{cases} \quad (3.2.14)$$

Observació 3.2.2. Utilitzant que $\dot{V} = \ddot{X}$ obtenim una EDO de segon ordre equivalent.

$$\ddot{X} + 2\beta' \dot{X} = -(\gamma' \text{Attn}(X) + \gamma'' \text{MLP}(X)) \quad (3.2.15)$$

Destaquem també que en passar al segon ordre l'estructura de [4] , es perd el paràmetre μ' relacionat amb el *lookahead*, com que μ' sempre multiplica termes d'ordre $\mathcal{O}(\Delta t^2)$ desapareix quan fem $\Delta t \rightarrow 0$. Llavors el desplaçament auxiliar de *symplectic-euler* queda sense efecte en fer el pas de l'esquema discret a la dinàmica contínua. Aquest desplaçament és el que marca la diferència entre l'acceleració de Nesterov i la que veurem al següent capítol (Polyak *heavy-ball*), en perdre-la, els dos límits continus coincideixen.

Capítol 4

Connexió amb *heavy-ball* i anàlisi de l'energia

A la secció anterior hem arribat a l'EDO

$$\ddot{X} + 2\beta' \dot{X} = -(\gamma' \text{Attn}(X) + \gamma'' \text{MLP}(X))$$

En aquest capítol la interpretarem com una variant del mètode d'optimització *heavy-ball* de [5] i estudiarem el comportament d'energia que té.

Recordem el context de [5]. Donada una funció a optimitzar $U : \mathbb{R}^m \rightarrow \mathbb{R}$ prou regular, el *heavy-ball* és un mètode que es representa com

$$\ddot{X} + \eta \dot{X} = -\nabla U(X) \quad (4.0.1)$$

amb $\eta > 0$ un coeficient de fregament. Ho podem veure com que $X(t)$ modela la trajectòria d'una bola amb inèrcia que es mou sobre la superfície de U . El sistema està impulsat pel gradient negatiu cap a mínims locals (valls) i a la vegada es va frenant degut al fregament η , proporcional a la velocitat $\dot{X}$. Aquesta fricció provoca que l'energia total del sistema es vagi dissipant fins que la bola s'estabilitza en un punt d'equilibri local.

Per ser coherents amb com hem definit el potencial anteriorment hem d'imposar $\gamma' = \gamma'' = 1$ per preservar

$$\nabla U(X) = \text{Attn}(X) + \text{MLP}(X) \quad \text{per a tot } X \in \mathbb{R}^{n \times d} \quad (4.0.2)$$

Amb aquesta assignació tenim

$$\begin{cases} \dot{X} = V \\ \dot{V} = -2\beta' V - (\text{Attn}(X) + \text{MLP}(X)) \end{cases} \quad (4.0.3)$$

Equivalentment obtenim l'EDO de segon ordre

$$\ddot{X} + 2\beta' \dot{X} = -(\text{Attn}(X) + \text{MLP}(X)) \quad (4.0.4)$$

Ara ens fixem en la similitud de les últimes EDO de segon ordre.

Proposició 4.0.1. *Suposant que existeix $U : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}$ amb $\nabla U(X) = \text{Attn}(X) + \text{MLP}(X)$, llavors l'equació (4.0.4) coincideix amb (4.0.1) prenent $2\beta'$ com a fregament.*

Demostració. Juntant (4.0.4) i que per (4.0.2) tenim:

$$-\nabla U(X) = -(\text{Attn}(X) + \text{MLP}(X)) \quad (4.0.5)$$

veiem que és equivalent a (4.0.1) amb $\eta = 2\beta'$. $\square$

La Proposició 4.0.1 ens fa pensar en el cas particular on $\beta' = 0$, desapareix la fricció. Per facilitar els càlculs al llarg del treball utilitzarem el producte de Frobenius per a matrius $A, B \in \mathbb{R}^{n \times d}$, on $A \cdot B := \sum_{i,l} A_{il} B_{il}$

Proposició 4.0.2. *Si $\beta' = 0$, l'energia $H(X, V) = \frac{1}{2}\|V\|^2 + U(X)$ és constant al llarg de les trajectòries del sistema (4.0.3).*

Demostració. Per la regla de la cadena,

$$\frac{d}{dt}H = \nabla_X H \cdot \dot{X} + \nabla_V H \cdot \dot{V} = \nabla U(X) \cdot V + V \cdot \dot{V}$$

Amb $\beta' = 0$ el sistema (4.0.3) és $\dot{X} = V$ i $\dot{V} = -\nabla U(X)$ i llavors tenim

$$\frac{d}{dt}H = \nabla U(X) \cdot V + V \cdot (-\nabla U(X)) = 0$$

$\square$

Ara anem a explorar com és el comportament quan $\beta' > 0$. No esperem que H es conservi però sí que decreixi al llarg del temps ja que estem en el cas dissipatiu, amb el següent resultat formalitzem la nostra intuïció, això és molt important pel control de l'estabilitat del sistema.

Teorema 4.0.3 (Decreixement de l'energia). *Segui $H(X, V) = \frac{1}{2}\|V\|^2 + U(X)$ amb $\nabla U = (\text{Attn} + \text{MLP})$ i $(X(t), V(t))$ una solució de (4.0.3). Llavors*

$$\frac{d}{dt}H(X(t), V(t)) = -2\beta' \|V(t)\|^2 \leq 0 \quad (4.0.6)$$

I H és monòtona no creixent.

Demostració. Per la regla de la cadena tenim

$$\frac{d}{dt}H = \nabla_X H \cdot \dot{X} + \nabla_V H \cdot \dot{V} = \nabla U(X) \cdot V + V \cdot \dot{V}$$

Ara substituint $\dot{V}$ de (4.0.3) i fent servir que $\nabla U = (\text{Attn} + \text{MLP})$,

$$\begin{aligned} V \cdot \dot{V} &= V \cdot [-2\beta' V - (\text{Attn}(X) + \text{MLP}(X))] \\ &= -2\beta' \|V\|^2 - (\text{Attn}(X) + \text{MLP}(X)) \cdot V \\ &= -2\beta' \|V\|^2 - \nabla U(X) \cdot V \end{aligned}$$

Sumant amb $\nabla U(X) \cdot V$ cancel·lem el segon terme i tenim

$$\frac{d}{dt}H = \nabla U \cdot V + (-2\beta'\|V\|^2 - \nabla U \cdot V) = -2\beta'\|V\|^2$$

Com $\beta' \geq 0$ obtenim que $\frac{d}{dt}H \leq 0$ que és el que buscàvem. $\square$

Pel següent lema cal recordar el Teorema de Weierstrass, la demostració és pot trobar a [13].

Teorema 4.0.4 (Weierstrass). *Sigui $K \subset \mathbb{R}^n$ compacte i $f : K \rightarrow \mathbb{R}$ contínua. Existeixen $x_{\min} \in K$ i $x_{\max} \in K$ tals que $f(x_{\min}) \leq f(x) \leq f(x_{\max})$, per a tot $x \in K$ (s'assoleixen el màxim i el mínim).*

Lema 4.0.5. *Assumint que existeix $U : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}$ amb $\nabla U(X) = \text{Attn}(X) + \text{MLP}(X)$, i suposant U és coerciva. Llavors, per a qualsevol condició inicial $(X_0, V_0) \in \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times d}$ existeix una única solució $(X(t), V(t))$ del sistema (4.0.3) definida $\forall t \in [0, \infty)$ i de classe $\mathcal{C}^1$.*

Demostració. Com tenim Attn , MLP , $\nabla U \in \mathcal{C}^1$ (per suposicions fetes anteriorment), els membres drets de les equacions del sistema 4.0.3 són $\mathcal{C}^1$ i també localment Lipschitz. Per [6, Teorema 2.2] existeix una única solució local en un entorn de $t = 0$, i per [6, Teorema 2.13] aquesta s'estén a una única solució maximal $(X(t), V(t))$ de classe $\mathcal{C}^1$ definida en un interval maximal (T_-, T_+) amb $T_- < 0 < T_+ \leq +\infty$. Com que la condició inicial és a $t = 0$ considerem la restricció a $[0, T_+)$. Volem veure que l'interval és $[0, \infty)$. Pel Teorema 4.0.3

$$\frac{d}{dt}H(X(t), V(t)) = -2\beta'\|V(t)\|^2 \leq 0 \quad \text{per a tot } t \in [0, T_+)$$

i per tant $H(X(t), V(t)) \leq H_0$. Amb això tenim:

- Com $\frac{1}{2}\|V(t)\|^2 \geq 0$, $U(X(t)) \leq H_0 \forall t$. Per la coercivitat de U , pel nivell H_0 existeix $R > 0$ tal que $U(X) > H_0$ sempre que $\|X\| > R$. Com $U(X(t)) \leq H_0$, necessitem $\|X(t)\| \leq R \forall t \in [0, T_+)$
- Sigui $m := \min_{\|X\| \leq R} U(X)$ que existeix pel Teorema de Weierstrass ja que U és contínua i $\{\|X\| \leq R\}$ compacte. Llavors

$$\frac{1}{2}\|V(t)\|^2 = H(X(t), V(t)) - U(X(t)) \leq H_0 - m,$$

d'on $\|V(t)\|^2 \leq 2(H_0 - m)$ per a tot $t \in [0, T_+)$.

Amb això tenim que la trajectòria està continguda en el compacte

$$K := \{\|X\| \leq R\} \times \{\|V\|^2 \leq 2(H_0 - m)\} \subset \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times d}$$

$\forall t \in [0, T_+)$. Si $T_+ < +\infty$ pel [6, Corol·lari 2.16] la solució ha d'estar fora de K quan $t \rightarrow T_+$, contradient que $(X(t), V(t)) \in K$. Per tant $T_+ = +\infty$. $\square$

Corol·lari 4.0.6. *Sota les hipòtesis del Lema 4.0.5, la trajectòria $(X(t), V(t))$ està fitada en $t \in [0, +\infty)$. Existeixen $R > 0$ i $M > 0$ tals que $\|X(t)\| \leq R$ i $\|V(t)\| \leq M$ per a tot $t \geq 0$.*

Demostració. Es dedueix de la demostració del Lema 4.0.5, on hem demostrat que la trajectòria està continguda al compacte $K = \{\|X\| \leq R\} \times \{\|V\|^2 \leq 2(H_0 - m)\}$
□

El Corol·lari 4.0.6 és dels resultats més importants que hem vist fins ara. Resol el problema que vèiem sobre l'explosió de les trajectòries al model de primer ordre.

Per avançar en l'anàlisi qualitatiu del sistema identificarem els punts d'equilibri.

Proposició 4.0.7 (Punts d'equilibri). *Un punt $(X, V) \in \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times d}$ és d'equilibri del sistema (4.0.3) si i només si*

$$V = 0 \quad i \quad \text{Attn}(X) + \text{MLP}(X) = 0 \quad (4.0.7)$$

Demostració. Un punt és d'equilibri si i només si es compleixen $\dot{X} = 0$ i $\dot{V} = 0$. De $\dot{X} = V = 0$ obtenim $V = 0$. Substituint a la segona equació,

$$0 = -2\beta' \cdot 0 - (\text{Attn}(X) + \text{MLP}(X)) = -(\text{Attn}(X) + \text{MLP}(X))$$

Equivalentment $\nabla U(X) = \text{Attn}(X) + \text{MLP}(X) = 0$ □

Corol·lari 4.0.8. *Suposem que existeix $U : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}$ amb $\nabla U(X) = \text{Attn}(X) + \text{MLP}(X)$ i $\beta' > 0$ i U coerciva per a qualsevol condició inicial (X_0, V_0) la solució de (4.0.3) satisfà*

$$\int_0^\infty \|V(t)\|^2 dt \leq \frac{H_0 - m}{2\beta'} < +\infty$$

on m és el mínim de U sobre el compacte $\{\|X\| \leq R\}$ de la demostració del Corol·lari 4.0.6.

Demostració. Pel Teorema 4.0.3 sabem que $\forall t \geq 0$,

$$\frac{d}{dt} H(X(t), V(t)) = -2\beta' \|V(t)\|^2$$

La funció $t \mapsto H(X(t), V(t))$ és $\mathcal{C}^1$ en $[0, \infty)$ ja que H és $\mathcal{C}^2$ i $(X(t), V(t))$ és $\mathcal{C}^1$ com a solució del sistema, integrem la identitat anterior entre 0 i T pel teorema fonamental del càlcul,

$$H(X(T), V(T)) - H_0 = -2\beta' \int_0^T \|V(t)\|^2 dt$$

o equivalentment,

$$2\beta' \int_0^T \|V(t)\|^2 dt = H_0 - H(X(T), V(T)) \quad (4.0.8)$$

Pel Corol·lari 4.0.6 la trajectòria es manté en el compacte $\{\|X\| \leq R\}$, sobre el qual U assoleix un mínim m pel teorema de Weierstrass. Degut a $\frac{1}{2}\|V(T)\|^2 \geq 0$,

$$H(X(T), V(T)) = \frac{1}{2}\|V(T)\|^2 + U(X(T)) \geq m$$

Substituint a (4.0.8),

$$\int_0^T \|V(t)\|^2 dt = \frac{H_0 - H(X(T), V(T))}{2\beta'} \leq \frac{H_0 - m}{2\beta'} \quad \forall T > 0.$$

Per tant hem trobat una cota.

Ara farem el pas al límit $T \rightarrow \infty$. La funció $F(T) := \int_0^T \|V(t)\|^2 dt$ és monòtona creixent i l'integrand és positiu, cap de les dues estrictament, com hem vist que està fitada superiorment per $\frac{H_0 - m}{2\beta'}$ existeix el límit

$$\int_0^\infty \|V(t)\|^2 dt = \lim_{T \rightarrow \infty} F(T) \leq \frac{H_0 - m}{2\beta'}$$

□

Ara que tenim els punts d'equilibri del sistema 4.0.3, que són de la forma $(X^*, 0)$ amb X^* punt crític del potencial, és a dir que $\nabla U(X^*) = 0$. Anem a fixar-nos com són des del punt de vista dinàmic. Si variem la condició inicial al voltant d'alguna trajectòria pot tornar-hi (equilibri estable), convergir-hi (asimptòticament estable) o allunyar-se'n (inestable). Formalment,

- $(X^*, 0)$ és *estable* si $\forall \varepsilon > 0$ existeix $\delta > 0$ tal que tota solució $(X(t), V(t))$ amb $\|(X(0), V(0)) - (X^*, 0)\| < \delta$ satisfà $\|(X(t), V(t)) - (X^*, 0)\| < \varepsilon$ per tot $t \geq 0$.
- És *asimptòticament estable* si és estable i existeix $\delta > 0$ on tota solució amb $\|(X(0), V(0)) - (X^*, 0)\| < \delta$ compleix $\lim_{t \rightarrow \infty} (X(t), V(t)) = (X^*, 0)$.
- És *inestable* si no és estable

Ara classificarem els punts entre aquests 3 tipus de comportaments segons la Hessiana $\nabla^2 U(X^*)$ i el coeficient de fricció β' .

Primer de tot hem de calcular els valors propis en el sistema linealitzat al voltant de l'equilibri. Enunciar aquest teorema que utilitzarem durant la demostració, la prova és pot trobar a [14], per exemple.

Teorema 4.0.9 (Teorema espectral). *Sigui $A \in \mathbb{R}^{n \times n}$ simètrica, llavors existeix una base ortonormal de $\mathbb{R}^n$ formada per vectors propis de A , on tots els valors propis són reals.*

Proposició 4.0.10. *Sigui $(X^*, 0)$ un equilibri del sistema (4.0.3) i $M := \nabla^2 U(X^*) \in \mathbb{R}^{nd \times nd}$. Llavors el sistema linealitzat al voltant de $(X^*, 0)$ és $\dot{Z} = AZ$ amb*

$$A = \begin{pmatrix} 0 & I_{nd} \\ -M & -2\beta' I_{nd} \end{pmatrix} \in \mathbb{R}^{2nd \times 2nd}$$

i els seus $2nd$ valors propis (comptats amb multiplicitat) venen donats per

$$\lambda_i^\pm = -\beta' \pm \sqrt{(\beta')^2 - \mu_i}, \quad i = 1, \dots, nd \quad (4.0.9)$$

on $\mu_1, \dots, \mu_{nd}$ són els valors propis de M .

Demostració. El sistema (4.0.3) s'escriu com $\dot{Z} = F(Z)$ amb $Z = (X, V)$ i

$$F(X, V) = (V, -2\beta'V - \nabla U(X))$$

El jacobiana de F és

$$DF(X, V) = \begin{pmatrix} 0 & I_{nd} \\ -\nabla^2 U(X) & -2\beta' I_{nd} \end{pmatrix}$$

avaluant-ho a $(X^*, 0)$ s'obté A .

Utilitzem $\mathbb{R}^{n \times d} \cong \mathbb{R}^{nd}$ per aplanar les matrius com a vectors per tenir $M \in \mathbb{R}^{nd \times nd}$ i poder calcular els valors propis. La matriu A té dimensions $2nd \times 2nd$, per tant no és adient calcular el polinomi característic utilitzant el determinant. Ens fixem en l'estructura per blocs, 0 , I_{nd} , $-M$ i $-2\beta' I_{nd}$. Veiem que M , és una matriu $nd \times nd$ que podem diagonalitzar utilitzant el teorema espectral.

Sigui (X, V) un vector propi de A amb valor propi $\lambda \in \mathbb{C}$, on $X, V \in \mathbb{R}^{nd}$. L'equació $AZ = \lambda Z$ es desenvolupa com

$$\begin{pmatrix} 0 & I_{nd} \\ -M & -2\beta' I_{nd} \end{pmatrix} \begin{pmatrix} X \\ V \end{pmatrix} = \lambda \begin{pmatrix} X \\ V \end{pmatrix}$$

i obtenim el sistema

$$\begin{cases} V = \lambda X, \\ -MX - 2\beta'V = \lambda V \end{cases} \quad (4.0.10)$$

Ara substituint $V = \lambda X$ a la segona equació

$$-MX - 2\beta'\lambda X = \lambda^2 X$$

i reordenant

$$MX = -(\lambda^2 + 2\beta'\lambda)X \quad (4.0.11)$$

Fixant-nos en l'equació (4.0.11) notem que equival a la condició que X sigui vector propi de M amb valor propi $-(\lambda^2 + 2\beta'\lambda)$. Ara utilitzem el teorema espectral, com que M és simètrica (pel Teorema de Schwarz) existeix una base ortonormal $\{e_1, \dots, e_{nd}\}$ de $\mathbb{R}^{nd}$ formada per vectors propis de M i amb valors propis reals $\mu_1, \dots, \mu_{nd}$. Amb això deduïm que X ha de ser múltiple d'algun e_i . Existeixen $i \in \{1, \dots, nd\}$ i $a \in \mathbb{R} \setminus \{0\}$ tals que $X = ae_i$, amb

$$\mu_i = -(\lambda^2 + 2\beta'\lambda) \quad (4.0.12)$$

Per la primera equació de (4.0.10), $V = \lambda X = (\lambda a)e_i$ també és múltiple de e_i .

La condició (4.0.12) reordenada és

$$\lambda^2 + 2\beta'\lambda + \mu_i = 0$$

una equació de segon grau en λ amb solucions

$$\lambda_i^\pm = -\beta' \pm \sqrt{(\beta')^2 - \mu_i}$$

Per a cada $i \in \{1, \dots, nd\}$ obtenim doncs dos valors propis de A (o un amb multiplicitat 2), que sumen $2nd$ en total. Com que el polinomi característic de A té grau $2nd$ hem trobat tots els valors propis que buscàvem. $\square$

Ara classificarem els punts d'equilibri amb el següent teorema.

Teorema 4.0.11. *Si $(X^*, 0)$ un equilibri amb $\nabla^2 U(X^*)$ no singular, llavors*

1. $(X^*, 0)$ és estable $\iff \nabla^2 U(X^*) \succ 0$.
2. $(X^*, 0)$ és asimptòticament estable $\iff \nabla^2 U(X^*) \succ 0$ i $\beta' > 0$.
3. $(X^*, 0)$ és inestable $\iff \nabla^2 U(X^*)$ té algun valor propi estrictament negatiu.

Demostració. Siguin $\mu_1, \dots, \mu_{nd}$ els valors propis de $\nabla^2 U(X^*)$. Com estem en el cas no degenerat, tenim $\mu_i \neq 0 \forall i$, i podem classificar els equilibris segons si:

A tots els $\mu_i > 0$ (Hessiana definida positiva),

B tots els $\mu_i < 0$ (Hessiana definida negativa),

C hi ha $\mu_i > 0$ i $\mu_j < 0$ per a algun i, j .

Per una banda demostrarem que (A) és estable. I també asimptòticament estable si i només si $\beta' > 0$, i per l'altra banda que tant (B) com (C) són inestables.

Comencem pel cas A. Com $\nabla U(X^*) = 0$ i $\nabla^2 U(X^*) \succ 0$ pel teorema de Taylor a segon ordre tenim que X^* és un mínim local de U estrictament. Pel següent pas necessitem veure que la funció W és una funció de Lyapunov local.

$$W(X, V) := H(X, V) - H(X^*, 0) = \frac{1}{2}\|V\|^2 + U(X) - U(X^*)$$

Hem de comprovar tres condicions,

(i) $W(X^*, 0) = 0$ Aixó es compleix per com hem construït la funció.

$$W(X^*, 0) = \frac{1}{2}\|0\|^2 + U(X^*) - U(X^*) = 0$$

(ii) W és definida positiva en un entorn de $(X^*, 0)$.

Per a $(X, V) \neq (X^*, 0)$ a l'entorn tenim que si $V \neq 0$ llavors $\frac{1}{2}\|V\|^2 > 0$ i $U(X) - U(X^*) \geq 0$ perquè és un mínim local, llavors $W > 0$.

Si $V = 0$ tenim que $X \neq X^*$, llavors $\frac{1}{2}\|V\|^2 = 0$ i $U(X) - U(X^*) > 0$ ja que és un mínim estricta, concloent amb $W > 0$.

(iii) $\frac{d}{dt}W(X(t), V(t)) \leq 0$. Utilitzant el Teorema 4.0.3 tenim

$$\frac{d}{dt}W(X(t), V(t)) = \frac{d}{dt}H(X(t), V(t)) = -2\beta'\|V(t)\|^2 \leq 0$$

Aplicant el teorema de Lyapunov [6, Teorema 6.13], $(X^*, 0)$ és estable.

Veiem que és asimptòtica si i només si $\beta' > 0$. Per la Proposició 4.0.10 amb $\mu_i > 0 \forall i$, els valors propis del sistema són $\lambda_i^\pm = -\beta' \pm \sqrt{(\beta')^2 - \mu_i}$, explorem-ho segons el valor de β' .

Si $\beta' > 0$ anem a veure que tots els $\lambda_i^\pm$ tenen part real estrictament negativa. Distingim dos casos segons el signe de $(\beta')^2 - \mu_i$.

- Si $(\beta')^2 < \mu_i$

$\sqrt{(\beta')^2 - \mu_i} = i\sqrt{\mu_i - (\beta')^2}$ és imaginari pur i

$$\lambda_i^\pm = -\beta' \pm i\sqrt{\mu_i - (\beta')^2}, \quad \text{Re}(\lambda_i^\pm) = -\beta' < 0$$

- Si $(\beta')^2 \geq \mu_i$

$\sqrt{(\beta')^2 - \mu_i}$ és real i compleix $0 \leq \sqrt{(\beta')^2 - \mu_i} < \sqrt{(\beta')^2} = \beta'$ ja que $\mu_i > 0$. Per tant

$$\lambda_i^- = -\beta' - \sqrt{(\beta')^2 - \mu_i} < 0, \quad \lambda_i^+ = -\beta' + \sqrt{(\beta')^2 - \mu_i} < 0$$

En els dos casos $\text{Re}(\lambda_i^\pm) < 0$. i pel teorema de linealització de Lyapunov [6, Teorema 6.10] l'equilibri és exponencialment estable i per tant asimptòticament estable.

Anem a veure ara que passa si $\beta' = 0$.

Els valors propis són $\lambda_i^\pm = \pm i\sqrt{\mu_i}$ imaginaris purs. Per la proposició 4.0.2, H és constant al llarg de les trajectòries i per tant $(X(t), V(t))$ està a la superfície de nivell $\{H = H_0\}$. Com $H(X^*, 0) = 0 + U(X^*) = U(X^*)$, $(X^*, 0)$ no pot formar part d'aquest conjunt de nivell a menys que $H_0 = H(X^*, 0)$. Si la condició inicial està en un entorn de $(X^*, 0)$ i $H_0 > U(X^*)$, la trajectòria mai convergirà. Per tant l'equilibri no és asimptòticament estable.

Ara tractem el cas B on $\nabla^2 U(X^*) \prec 0$.

Com tots els $\mu_i < 0$ tenim $\sqrt{(\beta')^2 - \mu_i} = \sqrt{(\beta')^2 + |\mu_i|} > \beta' \geq 0$, d'on

$$\lambda_i^+ = -\beta' + \sqrt{(\beta')^2 + |\mu_i|} > 0$$

Llavors tots els valors propis λ_i^+ tenen part real positiva i utilitzant el mètode indirecte de Lyapunov [7, Teorema 4.7] l'equilibri és inestable.

Tractem l'últim cas C amb $\nabla^2 U(X^*)$ indefinida.

Com existeixen i amb $\mu_i < 0$ podem seguir el mateix raonament del cas anterior i tenim $\lambda_i^+ > 0$, aplicant el mateix [7, Teorema 4.7(2)] veiem que l'equilibri és inestable.

Amb això hem demostrat l'enunciat. $\square$

Observació 4.0.12. No cobrim el cas on $\nabla^2 U(X^*)$ és singular perquè requereix tècniques d'ordre superior que no abordem en aquest treball.

Corol·lari 4.0.13. Si $\beta' > 0$ el sistema (4.0.3) no admet trajectòries periòdiques que no siguin constants.

Demostració. Suposem per arribar a una contradicció que $(X(t), V(t))$ és una trajectòria de període $T > 0$, és a dir, $(X(t+T), V(t+T)) = (X(t), V(t)) \forall t \geq 0$. Llavors tenim $H(X(T), V(T)) = H(X(0), V(0))$ i pel Teorema 4.0.3, $\frac{d}{dt}H(X(t), V(t)) = -2\beta'\|V(t)\|^2$. Integrant entre 0 i T

$$H(X(T), V(T)) - H(X(0), V(0)) = -2\beta' \int_0^T \|V(t)\|^2 dt = 0$$

Com que $\beta' > 0$, l'única manera que la integral sigui zero és que $\|V(t)\|^2 = 0 \forall t \in [0, T]$, però això implica $\dot{X} = V = 0$ i X constant, fent que la trajectòria també ho sigui i per tant no és periòdica no constant. $\square$

Observació 4.0.14. Aquest resultat és força intuïtiu ja que la fricció redueix energia al sistema i una trajectòria no pot tornar (a menys que sigui una trajectòria constant) a cap estat anterior perquè fer-ho significaria que té la mateixa energia que abans.

Capítol 5

Simulacions numèriques

Al llarg del treball hem anat construint la relació entre l'esquema de [4, B.2.Symplectic-Euler] i el sistema de segon ordre que buscàvem, i també hem deduït diverses propietats com l'acotació de les trajectòries al Corol·lari 4.0.6 o la caracterització de punts d'equilibri del Teorema 4.0.11.

Algunes de les propietats les hem obtingut amb l'ajuda de les simulacions que veurem, que han servit per reflexionar sobre el comportament observat i formular noves proposicions o corol·laris que hem demostrat. En aquest sentit han estat una font d'idees pels resultats teòrics del treball. D'altra banda, el rol que tenen és verificar diversos resultats que hem obtingut fins ara. En cas de veure resultats diferents dels esperats, això pot significar que o bé hi ha error en la implementació o en la teoria.

També ens ajuden a mostrar fenòmens i diferents aspectes de les trajectòries que no podem veure directament de l'estudi teòric, com els intercanvis d'energia cinètica i potencial, oscil·lacions... Això ens ajuda a identificar preguntes noves i aspectes teòrics a explorar durant el treball o en el futur.

Començarem els experiments amb una configuració senzilla i validant aspectes elementals de la teoria per poder anar afegint-hi més complexitat (com poden ser més tokens, dimensions o diferents MLP). Fem això per poder detectar errors fàcilment a la implementació en cas que hi hagi sospites d'un comportament que no és coherent amb el que hem vist fins ara. Cada experiment anirà responnent a preguntes que hagin quedat obertes pels anteriors.

Per totes les simulacions d'aquest capítol hem utilitzat *Python* (*NumPy*, *CuPy*, *Matplotlib*, *Sklearn*, *Seaborn*) que s'han executat o bé de forma local o en plataformes com Google Colab per experiments més potents.

Per garantir que tots els experiments es poden reproduir quan generem condicions inicials aleatòries fixem el generador pseudoaleatori de *NumPy* a cada experiment.

Durant els experiments registrem a cada pas k l'energia total $H_k = \frac{1}{2}\|V_k\|^2 + U(X_k)$, l'energia cinètica, la potencial i les normes $\|x_i(t)\|$ de cada *token*.

Experiment 1: Comportament de l'energia en el cas conservatiu

Amb aquest experiment volem verificar aspectes bàsics del nostre sistema com la conservació de H (tenint en compte mini-fluctuacions) al llarg de la dinàmica com enuncia la Proposició 4.0.2. També veure que la implementació compleix $\nabla U(X) = \text{Attn}(X)$ i fer-nos una idea del comportament qualitatiu de les trajectòries.

Utilitzarem la configuració més senzilla possible, així construïm; una base sòlida per anar complicant l'experiment poc a poc. Detalladament

- Dimensió $d = 3$ i nombre de *tokens* $n = 2$.
- Com a posicions inicials prenem dos punts unitaris a S^2 separats per $\theta_0 = \pi/3$

$$X_0 = \begin{pmatrix} 1 & 0 & 0 \\ \cos(\pi/3) & \sin(\pi/3) & 0 \end{pmatrix}$$

- Per les velocitats inicials partim del repós: $V_0 = 0$.
- Com a paràmetres prenem $\mu' = 0$, $\gamma' = 1 = \gamma''$ i $\beta' = 0$ (cas conservatiu).
- Fixem la discretització a $\Delta t = 10^{-2}$, $T_{\text{final}} = 15$ i $N = 1500$ passos.
- Prenem $\text{Attn}(X) = \frac{1}{n} X X^T X$, $MLP \equiv 0$ i $U(X) = \frac{1}{4n} \|X X^T\|_F^2$ com a (3.1.10)

Abans de simular l'experiment ens parem per comprovar numèricament que es compleix $\nabla U = \text{Attn}$ generant 1000 matrius X aleatòries i comparant $\text{Attn}(X)$ amb el gradient de U aproximat mitjançant diferències finites centrades amb pas $\varepsilon = 10^{-6}$. L'error màxim que hem vist és

$$\max_X \|\nabla U(X) - \text{Attn}(X)\|_\infty < 1.1 \times 10^{-8}$$

Llavors confirmem que encaixa amb el que esperàvem.

Amb el experiment obtenim la Figura 5.1 on es resumeixen diferents magnituds al llarg de la simulació i a la Figura 5.2 les trajectòries en $\mathbb{R}^3$.

Mesura	Valor
$H_0 = U(X_0)$	0,3125
$\max_t H(t) - H_0 $	$1,55 \times 10^{-3}$
$\max_t H(t) - H_0 / H_0 $	$4,97 \times 10^{-3}$
$\min_t \ x_1(t)\ $	0,0704
$\max_t \ x_1(t)\ $	1,0000

Amb aquesta taula resumim diferents mesures molt importants, veiem que l'energia es conserva ja que la diferència relativa $\max_t |H(t) - H_0|/|H_0|$ és prou petita. I també veiem que les trajectòries estan acotades per el màxim i mínim de les normes. Per tant tot es comporta com esperavem.

Si ens fixem en la gràfica de la variació d'energia $H(t) - H_0$ (Figura 5.1, dalt esquerra) podem observar que oscil·la periòdicament al voltant de zero i l'amplitud màxima sembla constant en el temps.

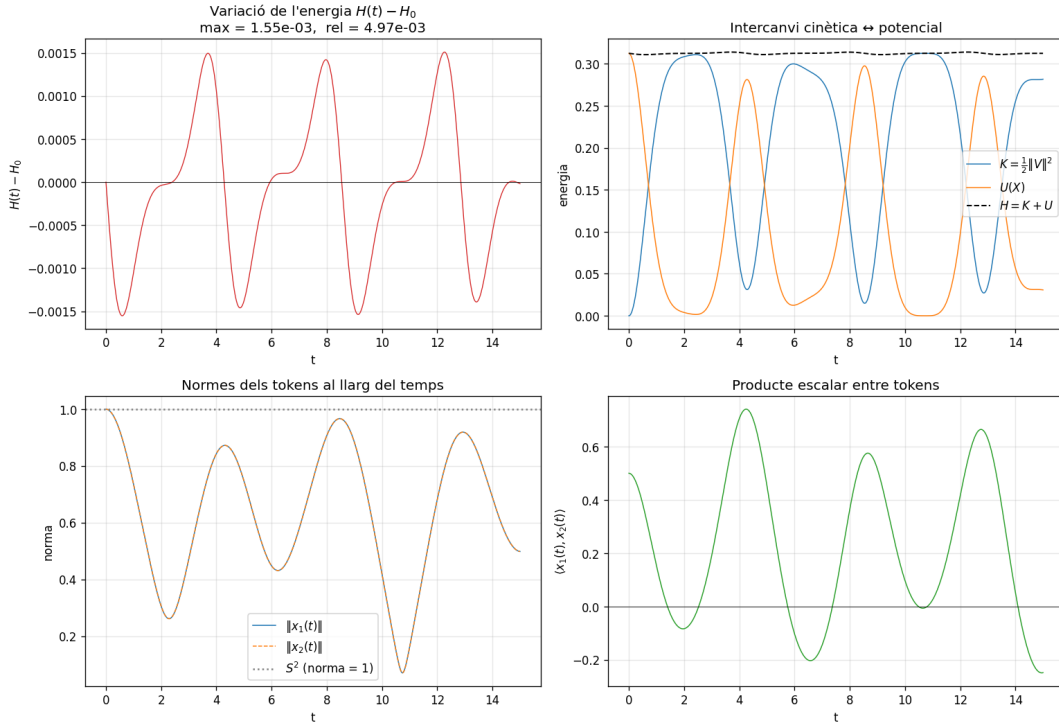


Figura 5.1: Gràfiques de l'Experiment 1. **Dalt esquerra:** Variació de l'energia $H(t) - H_0$ al llarg del temps. **Dalt dreta:** Intercanvi de l'energia cinètica i potencial. **Baix esquerra:** Normes dels *tokens*. **Baix dreta:** Producte escalar $\langle x_1(t), x_2(t) \rangle$.

A la gràfica de les energies cinètica i potencial s'aprecia un comportament oscil·latori molt regular, K i U s'intercanvien i a la vegada la suma és casi constant. Els *tokens* arriben a $\|x_i\| \approx 0,07$ abans d'arribar al mínim del potencial a la gràfica i sortir disparats per la energia cinètica acumulada K .

També veiem que tot i partir de S^2 els tokens no es queden allà ja que recorren l'interval de normes $\|x_i\| \in [0,07, 1,00]$. Per tant, sense afegir-hi una projecció no podem preservar que es mantinguin en S^2 . Cal destacar que a implementacions reals de *transformers* això es controla utilitzant el *LayerNorm*, que no podem implementar perquè trencaria la condició $\nabla U = \text{Attn} + \text{MLP}$ sobre la qual hem construït el potencial.

També notem que les coordenades z de les trajectòries (Figura 5.2) són nul·les durant tot l'experiment perquè inicialment no tenen component z i el camp $\text{Attn}(X) = \frac{1}{n}XX^\top X$ tampoc la introdueix, fent que les trajectòries siguin 2D. Per estudiar dinàmiques que siguin 3D afegirem un token més al següent experiment, i cap de les components serà nula.

Ara que ja ahem estudiat el cas conservatiu amb $\beta' = 0$ ens preguntem com és comportaria tot amb $\beta' \geq 0$.

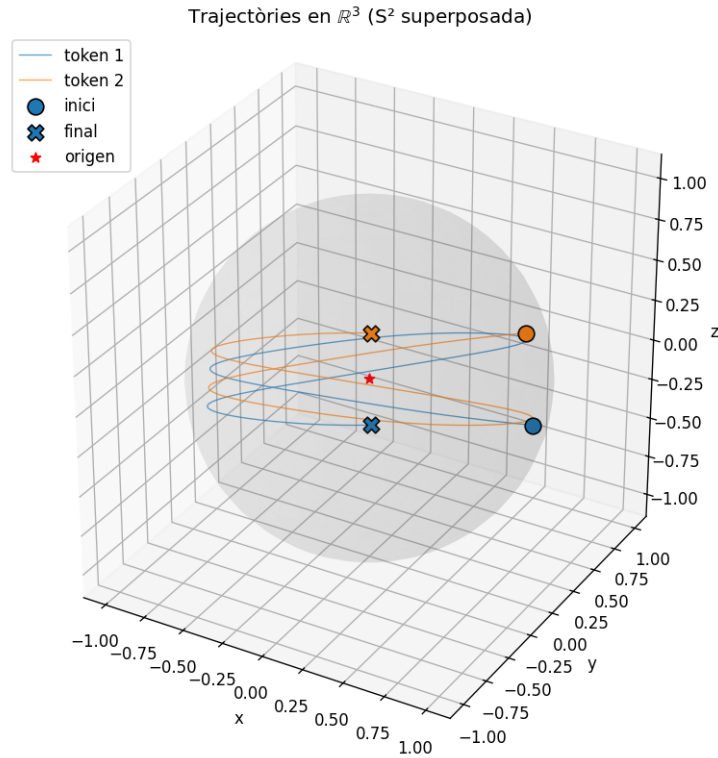


Figura 5.2: Trajectòries dels dos *tokens* en $\mathbb{R}^3$ al llarg de $T = 15$. La S^2 es dibuixa només com a referència visual, les trajectòries no la preserven.

Experiment 2: cas dissipatiu i convergència a equilibri

Ara estudiarem el cas dissipatiu i verifiquem diferents resultats del Capítol 4 com que $H(t)$ és monòtona no creixent pel Teorema 4.0.3. També comprovarem que $\dot{H}(t) = -2\beta'\|V(t)\|^2$, l'integral $\int_0^T \|V(t)\|^2 dt$ respecta la cota del Corol·lari 4.0.8 i que les trajectòries convergeixen a un punt d'equilibri ($V(t) \rightarrow 0$ i $\nabla U(X(t)) \rightarrow 0$) com indica la Proposició 4.0.7. Igual que a l'experiment anterior també ens fixarem en com evoluciona en general.

La nostra configuració aquest cop és:

- Dimensió $d = 3$ i nombre de *tokens* $n = 3$.
- Com a posicions inicials utilitzem tres vectors generats aleatòriament i projectats a S^2 amb `seed = 15`. Aquestes són:
 $x_0 = (-0.8153544; -0.533673; 0.2244779)$,
 $x_1 = (-0.47788402; 0.47927875; 0.73615131)$,
 $x_2 = (-0.7745936; 0.54575267; -0.31962286)$
 Amb velocitats inicials nul·les.
- Com a paràmetres fixem $\mu' = 0$, $\gamma' = \gamma'' = 1$ i $\beta' = 0, 1$.
- Per la discretització, $\Delta t = 5 \times 10^{-3}$, $T_{\text{final}} = 25$ i $N = 5000$ passos.

- Seguim mantenint $\text{Attn}(X) = \frac{1}{n}XX^TX$, $MLP \equiv 0$ i $U(X) = \frac{1}{4n}\|XX^T\|_F^2$ com a l'experiment anterior.

També executem en paral·lel la mateixa simulació amb $\beta' = 0$ per utilitzar-la com a control. Escollim $\beta' = 0,1$ perquè és prou petita per fer que la dinàmica conservi oscil·lacions visibles i la dissipació sigui visible a la vegada.

A la taula 5.1 trobem les mètriques principals de l'experiment. Podem confirmar que es compleixen els resultats que hem mencionat abans.

Mesura	Valor
H_0	0,303128
$H(T)$	$4,69 \times 10^{-4}$
Passos amb $\Delta H > 0$	0 de 5000
Error mitjà $ \dot{H} - (-2\beta'\ V\ ^2) $	$4,09 \times 10^{-5}$
Error màxim $ \dot{H} - (-2\beta'\ V\ ^2) $	$1,17 \times 10^{-3}$
$\int_0^T \ V\ ^2 dt$	1,512
Cota $\frac{H_0-m}{(2\beta')}$	1,516
Ratio integral/cota	0,998
$\ V(T)\ $	$2,99 \times 10^{-2}$
$\ \nabla U(X(T))\ $	$6,21 \times 10^{-4}$
$\ X(T)\ $	0,149

Taula 5.1: Resultats numèrics de l'Experiment 2.

Primer de tot, veiem que al llarg dels 5000 passos de la simulació no hi ha cap pas amb $\Delta H > 0$. Per tant $H(t)$ és no creixent al llarg de tota la trajectòria i compleix amb el Teorema 4.0.3. També veiem que l'error mitjà de $\dot{H} = -2\beta'\|V\|^2$ és $4,09 \times 10^{-5}$ i el màxim és $(1,17 \times 10^{-3})$. Aquests error són atribuïbles al mètode de diferenciació numèrica utilitzat.

També observem que la integral $\int_0^T \|V\|^2 dt = 1,512$ és molt pròxima a la cota teòrica $\frac{H_0-m}{2\beta'} = 1,516$ amb ratio de 0,998, per tant, gairebé tota l'energia inicial s'ha dissipat en només $T = 25$, això és coherent amb que $H(T) = 4,69 \times 10^{-4}$ és proper a zero. Si féssim el límit $T \rightarrow \infty$, esperaríem $H(\infty) = 0$ i arribaríem a la cota. A la figura 5.3 (Baix esquerra) veiem que la corba teòrica és indistingible de la numèrica.

A $t = T = 25$ les normes són $\|V\| = 2,99 \times 10^{-2}$, $\|\nabla U\| = 6,21 \times 10^{-4}$, i $\|X\| = 0,149$, les dues primeres són petites però no nul·les, per tant la trajectòria no ha arribat a l'origen encara, amb un β' prou gran la convergència seria més ràpida i podríem observar com $\|V\| = 0$ i $\|\nabla U\| = 0$.

A la Figura 5.5 podem veure clarament la diferència amb el cas $\beta' = 0$, en aquest les trajectòries oscil·len indefinidament sense convergir, en canvi al dissipatiu les espirals de les trajectòries es van apropant als punts d'equilibri.

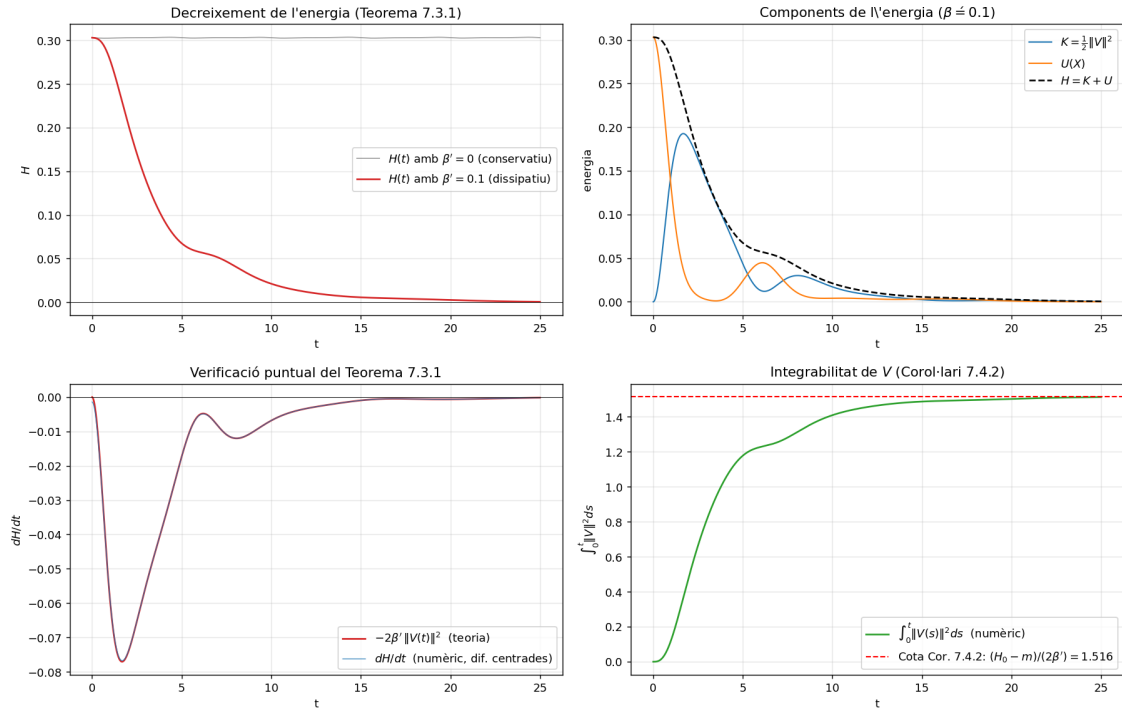


Figura 5.3: Gràfiques del experiment 2. **Dalt esquerra:** Decreixement monòton de $H(t)$ comparat el cas conservatiu. **Dalt dreta:** Energia total i intercanvi d'energia cinètica i potencial. **Baix esquerra:** Verificació de $\dot{H} = -2\beta'\|V\|^2$. **Baix dreta:** Comparació de $\int_0^t \|V\|^2 ds$ amb la cota $\frac{H_0 - m}{2\beta'}$.

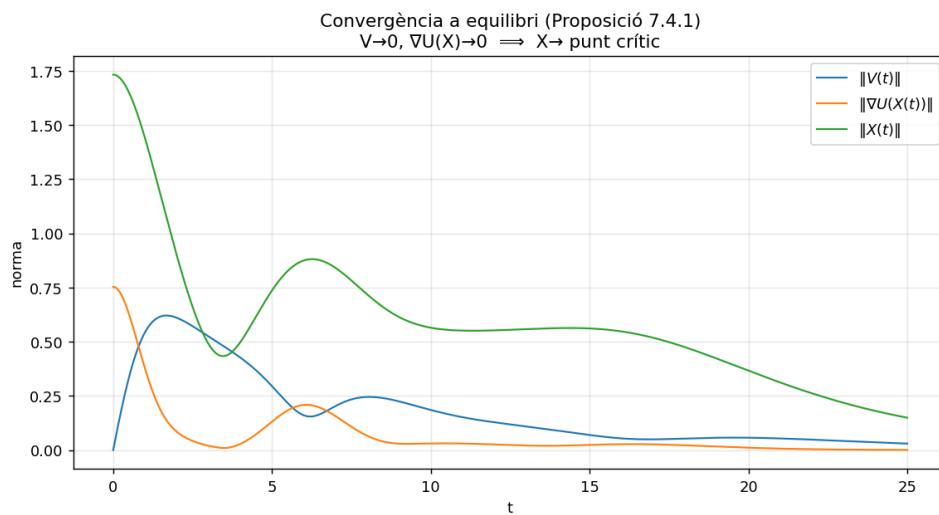


Figura 5.4: Evolució de $\|V(t)\|$, $\|\nabla U(X(t))\|$ i $\|X(t)\|$.

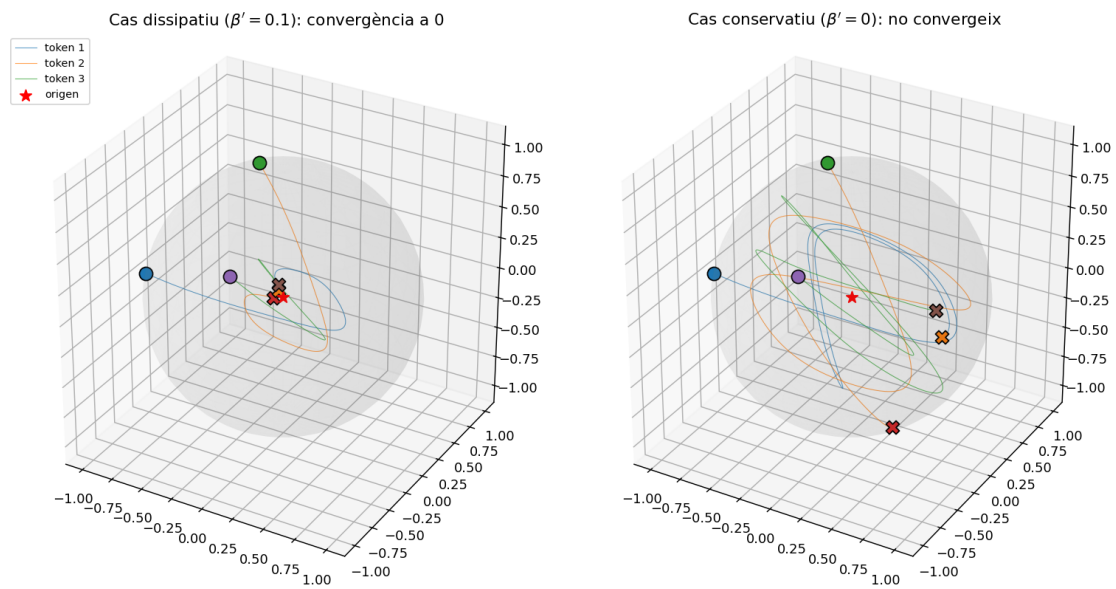


Figura 5.5: Trajectòries dels tres *tokens* en $\mathbb{R}^3$. **Esquerra:** Cas dissipatiu. **Dreta:** Cas conservatiu.

Experiment 3: caracterització de l'estabilitat dels equilibris

En aquest experiment volem centrarnos en la caracterització local dels punts d'equilibri, donada pel Teorema 4.0.11. Volem verificar numèricament les diferents parts que hem vist del teorema construint sistemes amb diferents tipus d'equilibris.

Prenem com a potencial

$$U(X) = U_{Attention}(X) + U_{MLP}(X) = \frac{1}{4n} \|XX^T\|_F^2 + U_{MLP}(X),$$

on U_{MLP} és el mateix que (3.1.7), recordem que

$$MLP(X)_{ij} = -\alpha_i X_{ij}, \quad U_{MLP}(X) = -\sum_{i=1}^n \frac{\alpha_i}{2} \|x_i\|^2$$

Per generar els tres casos que es veuen al Teorema 4.0.11, necessitem poder controlar els valors propis de la Hessiana $\nabla^2 U(X^*)$ on X^* compleix que $\nabla U(X^*) = \text{Attn}(X^*) + \text{MLP}(X^*) = 0$, notem que $X^* = 0$ compleix això.

Anem a calcular $\nabla^2 U(0)$, primer notem que la part $U_{Attn}(X)$ és homògena de grau 4, per tant no contribueix a $\nabla^2 U(0)$ ja que la seva hessiana s'anul·la a l'origen. I llavors $\nabla^2 U(0) = \nabla^2 U_{MLP}(0)$. Necessitem calcular-ho, notem que $\frac{\text{MLP}(X)_{ij}}{\partial X_{pq}} = -\alpha_i \delta_{ip} \delta_{jq}$ i per tant

$$\nabla^2 U(0) = \nabla^2 U_{MLP} = \frac{\partial(-\alpha_i X_{ij})}{\partial X_{pq}} = -\alpha_i \delta_{ip} \delta_{jq}$$

és una matriu diagonal de valors propis $-\alpha_i$ amb multiplicitat d de la forma $\text{diag}(\underbrace{-\alpha_1, \dots, -\alpha_1}_d, \underbrace{-\alpha_2, \dots, -\alpha_2}_d, -\alpha_3, \dots)$

Ara que podem controlar els valors propis $\nabla^2 U(0)$ generem els tres casos del Teorema 4.0.11 ajustant els signes dels α_i .

- Pel cas A volem que tots els valors propis $(-\alpha_i)$ siguin positius (Hessiana definida positiva), per tant prenem $\alpha = (-0, 5, -0, 5, -0, 5)$ i tenim valors propis de $+0, 5$ de multiplicitat 9.
- Al cas B volem que siguin negatius (Hessiana definida negativa), prenem $\alpha = (+0, 5, +0, 5, +0, 5)$, obtenint valors propis de $-0, 5$ de multiplicitat 9.
- Al cas C volem tenir-ne positius i negatius (Hessiana indefinida), prenem $\alpha = (+0, 5, -0, 5, +0, 1)$ de manera que els valors propis són $\{-0, 5; -0, 1; +0, 5\}$, cadascun amb multiplicitat 3.

La nostra configuració per aquest experiment és:

- Dimensió $d = 3$ i nombre de *tokens* $n = 3$

- Prenem com a pertorbació inicial $X_0 = \varepsilon \xi$ amb $\varepsilon = 0,05$ i ξ una matriu aleatòria pels casos A i B, pel C la pertorbació s'alinea amb la direcció inestable (la coordenada on el valor propi és negatiu, $-0,5$) i les altres coordenades s'inicialitzen a zero per evidenciar la inestabilitat.
- Velocitat inicial $V_0 = 0$
- Fixem els paràmetres $\mu' = 0$, $\gamma' = \gamma'' = 1$ i $\beta' = 0,1$
- Per la discretització, $\Delta t = 5 \times 10^{-3}$, $T_{\text{final}} = 120$ i $N = 24\,000$ passos.

La següent taula resumeix els resultats numèrics dels tres casos.

Cas	Valors propis $\nabla^2 U(0)$	$\ X(0)\ $	$\ X(T)\ $	$\frac{\ X(T)\ }{\ X(0)\ }$
A	$\{+0,5\}$	0,144	$3,0 \times 10^{-4}$	$2,4 \times 10^{-3}$
B	$\{-0,5\}$	0,144	2,123	14,79
C	$\{-0,5, -0,1, +0,5\}$	0,073	1,224	16,87

Taula 5.2: Resultats numèrics dels tres casos del Teorema 4.0.11.

Amb aquests validem les tres parts del Teorema 4.0.11 ja que observem que la trajectòria del Cas A (estable) convergeix a l'origen, mentre que les dels casos B i C s'allunyen del punt d'equilibri. Explorem-los de forma més detallada.

Al cas A la norma $\|X(t)\|$ decreix de 0,144 a 3×10^{-4} això és una reducció en un factor de gairebé 500 i veiem que $H(t)$ s'aproxima a zero, coherent amb l'equilibri $X^* = 0$ on $U(0) = 0$. A la trajectòria observem oscil·lacions decreixents (gràfica dalt esquerra de la Figura 5.6) coherents amb una convergència oscil·latoria cap a l'equilibri. Per tant, el comportament estable és evident.

Al cas B la trajectòria s'allunya ràpidament de l'origen i s'estabilitza en $\|X\| \approx 2,12$. La inestabilitat de l'origen és prou clara, la trajectòria convergeix cap a un altre equilibri.

Al Cas C la trajectòria s'allunya per un factor 16,87 del seu valor inicial confirmant la part corresponent del Teorema 4.0.11.

A les Figures 5.7 (dreta) i 5.8 (dreta) podem observar la natura del punt de sella de la superfície del, veiem dues valls laterals (direccions estables) separades per una cresta que passa per l'origen (direcció inestable). Les dues trajectòries de la Figura 5.8 (dreta) parteixen d'una pertorbació similar però en direccions diferents, la que està a la direcció inestable s'allunya i la de la direcció estable es queda a prop de l'origen.

A tots el casos hem mesurat la distància $\|X(t) - X(0)\|$ respecte el temps (gràfica baix esquerra de la Figura 5.6) i notem que no s'observen òrbites periòdiques.

Experiment 4: exploració amb diferents formes del MLP

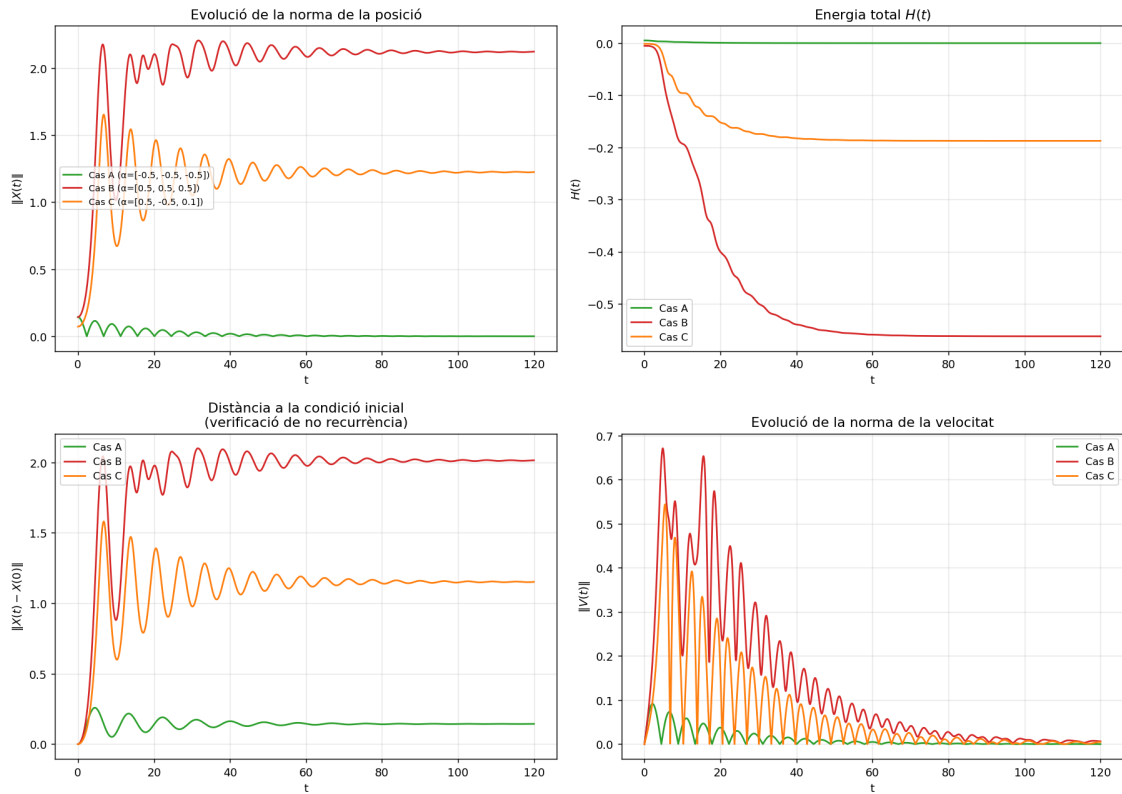


Figura 5.6: Verificació del Teorema 4.0.11. **Dalt esquerra:** Evolució de $\|X(t)\|$. **Dalt dreta:** Energia total $H(t)$. **Baix esquerra:** Distància a la condició inicial. **Baix dreta:** Evolució de $\|V(t)\|$.

En els experiments anteriors hem verificat diferents resultats teòrics treballant amb $d = 3$ i $n \leq 3$. Però els *Transformers* que s'utilitzen a la pràctica tenen un n molt més gran (centenars o milers de *tokens*) i embeddings d'alta dimensió. És natural preguntar-se com canviaria el comportament del sistema si estiguéssim en aquest cas. A més, diferents estudis com [11] demostren que la dinàmica de l'atenció *softmax* produeix *clustering* dels *tokens*. El que volem analitzar en aquest experiment és si el nostre el sistema té un comportament similar i com l'elecció del MLP afecta a aquest fenomen.

Cal destacar que aquest experiment té caràcter exploratori, no ens centrem en verificar cap teorema concret del treball, sinó que volem estudiar el comportament del sistema amb un nombre de tokens i dimensions superiors i MLP diferents. També visualitzem aquests casos amb dimensió $d = 3$.

Tindrem en compte diferents casos, sempre mantenint l'atenció lineal $\text{Attn}(X) = \frac{1}{n} X X^T X$. La taula 5.3 resumeix els vuit casos considerats.

Es comprova directament que cada MLP és gradient del potencial corresponent component a component i per tant seguim complint les suposicions que hem fet durant el treball. Notem que hem triat MLP de grau senar (X^3 i X^5) perquè generen potencials de grau parell (quàrtic, sèxtic) que preserven la coercivitat de U , ja que $\frac{\beta}{4} \sum X_{ij}^4$ i $\frac{\beta}{6} \sum X_{ij}^6$ són positius i dominants a l'infinit. Si prenem MLP

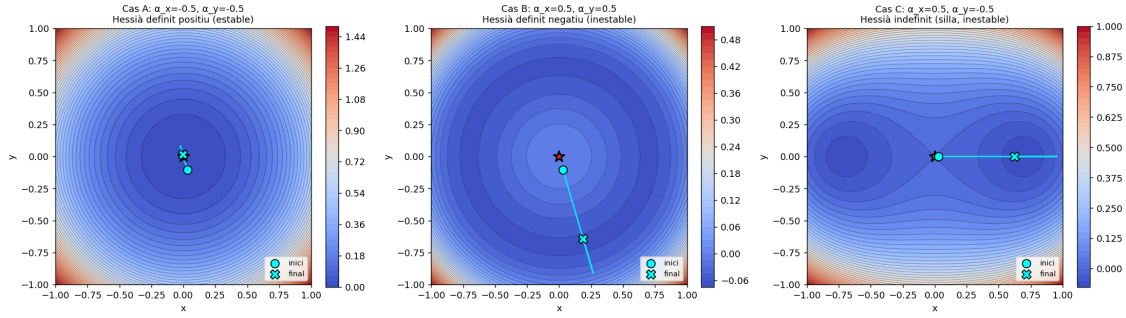


Figura 5.7: Visualització 2D de U_{total} amb $n = 1$ token en $\mathbb{R}^2$. **Esquerra:** Cas A (estable). **Centre:** Cas B (inestable). **Dreta:** Cas C (punt de sella).

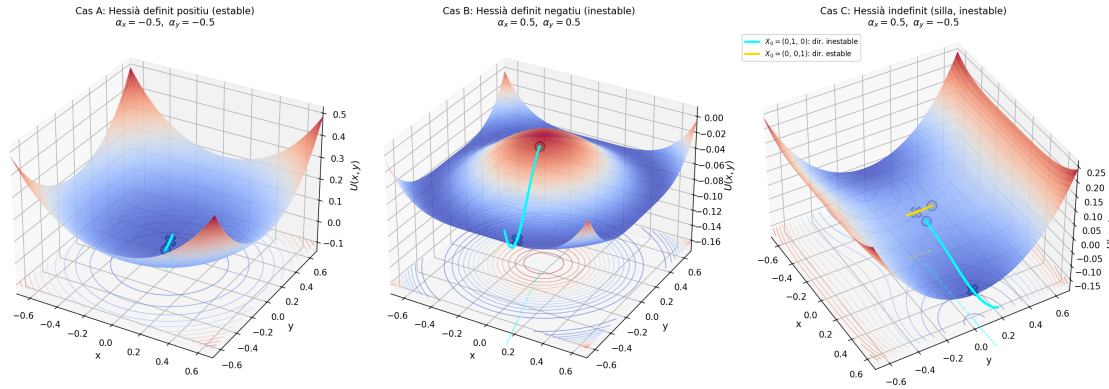


Figura 5.8: Visualització 3D de la superfície del potencial U per als tres casos. **Esquerra:** Cas A. **Centre:** Cas B. **Dreta:** Cas C. Es mostren *dues* trajectòries, la cyan parteix en la direcció inestable $(0, 1, 0)$, la groga parteix en la direcció estable $(0, 0, 1)$.

amb terme X^2 o X^4 faria que el potencial tinguí terme dominant X^3 o X^5 que podria tendir a $-\infty$ fent que no es satisfaci la hipòtesi de coercivitat.

Per a cada cas simulem la dinàmica amb set configuracions (n, d) diferents: $(100, 3)$; $(200, 4)$; $(300, 5)$; $(400, 8)$; $(800, 16)$; $(1600, 32)$ i $(3200, 64)$. Aquesta tria augmenta n i d amb una relació $n/d \approx 50$. No podem utilitzar configuracions amb n encara més gran o d molt superior perquè requeririen de GPUs més potents de les disponibles per a aquest treball. Llavors podem intuir que seguint amb la proporció mencionada, la densitat dels tokens anirà baixant conforme augmenta la dimensió d .

Per les visualitzacions dels estats finals i de la dinàmica utilitzem $(n, d) = (100, 3)$, per veure la evolució i intuir el comportament en dimensions superiors. Per a les gràfiques en les que volem observar el comportament en altes dimensions utilitzem $(n, d) = (3200, 64)$.

Els n tokens estan distribuïts de forma aleatoria a S^{d-1} amb $V_0 = 0$. Per la discretització utilitzem $\Delta t = 5 \times 10^{-3}$, temps final $T = 150$ i $N = 30000$ passos. Fixem els paràmetres $\mu' = 0$, $\gamma' = \gamma'' = 1$.

Cas	β'	MLP(X) $_{ij}$	Potencial U_{MLP}
C1: Conservatiu	0	-	0
C2: Dissipatiu sense MLP	0,1	-	0
C3: MLP lineal estable	0,1	$-\alpha X_{ij} \quad \alpha = -0,3$	$-\frac{\alpha}{2} \ X\ _F^2$
C4: MLP lineal inestable	0,1	$-\alpha X_{ij} \quad \alpha = +0,3$	$-\frac{\alpha}{2} \ X\ _F^2$
C5: MLP lineal indefinit	0,1	$-\alpha_i X_{ij} \quad \alpha \in \{-0,3; +0,3\}$	$-\sum_i \frac{\alpha_i}{2} \ x_i\ ^2$
C6: MLP cúbic repulsiu	0,1	$-\alpha X_{ij} + \beta X_{ij}^3 \quad \alpha = +0,3$	$-\frac{\alpha}{2} \ X\ _F^2 + \frac{\beta}{4} \sum X_{ij}^4$
C7: MLP cúbic indefinit	0,1	$-\alpha_i X_{ij} + \beta X_{ij}^3 \quad \alpha = +0,3$	$-\sum_i \frac{\alpha_i}{2} \ x_i\ ^2 + \beta X_{ij}^4$
C8: MLP quintic repulsiu	0,1	$-\alpha X_{ij} + \beta X_{ij}^5 \quad \alpha = +0,3$	$-\frac{\alpha}{2} \ X\ _F^2 + \frac{\beta}{6} \sum X_{ij}^6$

Taula 5.3: Els vuit casos explorats. Per C5 i C7 s'assignen $\alpha_i = -0,3$ a la meitat de les files i $\alpha_i = +0,3$ a l'altra meitat.

Hem triat $T = 150$ perquè els casos dissipatius s'aproximin suficientment a l'equilibri, com es veurà a partir de la velocitat final.

A la figura 5.9 veiem que les propietats demostrades durant el treball es mantenen vàlides en aquest cas d'alta dimensió, al cas C1, l'energia $H(t)$ es manté pràcticament constant, als casos dissipatius (C2–C8), $H(t)$ decreix monòtonament. Els teoremes es segueixen complint amb aquestes configuracions, tot i que els valors absoluts de H creixen molt en mòdul. Si ens fixem en la taula 5.4 veiem que $\|V(T)\|$ és petita en tots els casos dissipatius, de l'ordre de 10^{-3} a 10^{-2} a baixa dimensió i 10^{-1} a alta dimensió. Això ho interpretem com que estan convergint als diferents punts d'equilibri del sistema. El fet que $\|V(T)\|$ creixi amb la dimensió fa visible que quan més gran sigui d , el sistema necessita més temps per arribar a l'equilibri. En el cas C1 observem resultats esperables, l'energia cinètica s'intercanvia indefinidament amb la potencial sense apropar-se a zero.

La nomenclatura dels casos lineals (C3, C4, C5) l'escollim de la classificació del Teorema 4.0.11. A l'experiment anterior vam verificar aquesta classificació, a la taula 5.4 la veiem a una escala molt més gran.

Observem que al cas C3 (Hessiana definida positiva) l'origen és estable i les trajectòries convergeixen a $X = 0$, amb $\max_i \|x_i(T)\| \approx 10^{-3}$ a totes les configuracions. Per C4 (Hessiana definida negativa) l'origen és inestable i les trajectòries s'allunyen d'aquest, convergint a altres equilibris (en aquest cas, una esfera de radi ~ 1 a $d = 3$). Per C5 (Hessiana indefinida) l'origen és un punt de sella i les trajectòries s'allunyen al llarg de les direccions inestables (files amb $\alpha_i > 0$) i convergeixen al llarg de les estables ($\alpha_i < 0$), generant dues dinàmiques molt diferents. On una forma una esfera i l'altra es va concentrant a l'origen.

Fixant-nos en els casos no lineals (C6, C7, C8) tenen, a l'origen, que tenen la mateixa hessiana que C4 (per a C6, C8) o C5 (per a C7) i per tant l'origen és inestable. Notem que el comportament global és diferent, les trajectòries C6 i C8 es concentren cap a nous equilibris allunyats de l'origen, i C7 té el mateix comportament afegint un grup de *tokens* (exactament la meitat) que es concentra a l'origen. La formació d'agrupacions a C6, C7 i C8 sembla deguda a un major nombre d'equilibris provocada per la no linealitat. C1 i C2 es comporten com a

Cas	(100, 3)	(200, 4)	(300, 5)	(400, 8)	(800, 16)	(1600, 32)	(3200, 64)
$\max_i \ x_i(T)\ $							
C1	0,977	0,983	0,991	0,901	0,937	0,837	0,794
C2	0,040	0,039	0,046	0,046	0,062	0,105	0,085
C3	0,001	0,000	0,000	0,000	0,001	0,001	0,001
C4	1,032	1,186	1,306	1,676	2,337	3,273	4,570
C5	1,433	1,702	1,849	2,407	3,386	4,761	6,546
C6	0,928	1,080	1,200	1,509	2,152	3,021	4,245
C7	1,331	1,439	1,648	2,081	2,916	4,097	5,761
C8	0,959	1,153	1,240	1,559	2,211	3,113	4,377
$\text{diam}(X(T))$							
C1	1,939	1,952	1,966	1,677	1,732	1,485	1,326
C2	0,079	0,078	0,090	0,088	0,114	0,184	0,146
C3	0,001	0,001	0,000	0,001	0,001	0,001	0,001
C4	2,060	2,336	2,563	3,273	4,341	5,819	7,841
C5	2,814	3,325	3,627	4,495	6,109	8,267	11,109
C6	1,857	2,160	2,399	3,017	4,254	5,604	7,537
C7	2,662	2,877	3,296	4,114	5,606	7,516	10,086
C8	1,918	2,268	2,481	3,113	4,337	5,866	7,771
$\ V(T)\ $							
C1	3,34	3,95	3,81	4,39	3,68	4,15	4,33
C2	$1,2 \cdot 10^{-3}$	$2,2 \cdot 10^{-3}$	$2,4 \cdot 10^{-3}$	$1,4 \cdot 10^{-3}$	$4,0 \cdot 10^{-3}$	$1,3 \cdot 10^{-2}$	$6,7 \cdot 10^{-3}$
C3	$2,0 \cdot 10^{-3}$	$4,4 \cdot 10^{-3}$	$5,7 \cdot 10^{-3}$	$5,5 \cdot 10^{-3}$	$4,9 \cdot 10^{-3}$	$4,3 \cdot 10^{-3}$	$4,1 \cdot 10^{-3}$
C4	$1,4 \cdot 10^{-2}$	$2,4 \cdot 10^{-2}$	$3,3 \cdot 10^{-2}$	$5,0 \cdot 10^{-2}$	$1,0 \cdot 10^{-1}$	$2,1 \cdot 10^{-1}$	$3,9 \cdot 10^{-1}$
C5	$1,5 \cdot 10^{-2}$	$2,5 \cdot 10^{-2}$	$3,4 \cdot 10^{-2}$	$5,1 \cdot 10^{-2}$	$1,0 \cdot 10^{-1}$	$1,9 \cdot 10^{-1}$	$3,5 \cdot 10^{-1}$
C6	$1,2 \cdot 10^{-2}$	$2,0 \cdot 10^{-2}$	$2,8 \cdot 10^{-2}$	$4,1 \cdot 10^{-2}$	$8,3 \cdot 10^{-2}$	$1,7 \cdot 10^{-1}$	$3,4 \cdot 10^{-1}$
C7	$1,1 \cdot 10^{-2}$	$1,9 \cdot 10^{-2}$	$2,5 \cdot 10^{-2}$	$3,6 \cdot 10^{-2}$	$7,2 \cdot 10^{-2}$	$1,5 \cdot 10^{-1}$	$2,9 \cdot 10^{-1}$
C8	$1,7 \cdot 10^{-2}$	$2,2 \cdot 10^{-2}$	$3,1 \cdot 10^{-2}$	$4,5 \cdot 10^{-2}$	$9,2 \cdot 10^{-2}$	$2,0 \cdot 10^{-1}$	$4,0 \cdot 10^{-1}$

Taula 5.4: $\max_i \|x_i(T)\|$, $\text{diam}(X(T))$ i $\|V(T)\|$ a $T = 150$ per a les set configuracions simulades.

experiments anteriors.

Una observació interessant és que per C2 el potencial total és $U_{\text{Attn}}(X) = \frac{1}{4n} \|XX^T\|_F^2$ i la seva Hessiana a l'origen és zero, i per tant l'origen és un equilibri *degenerat* no cobert pel Teorema 4.0.11. Però podem veure que les trajectòries tendeixen lentament cap a l'origen, a $T = 150$ la norma màxima és $\sim 0,04$ a (100, 3) una mica més gran que a C3, sembla que la convergència existeix però és més lenta que en el cas no degenerat estable.

La taula 5.4 ens permet estudiar com els estats finals depenen de la dimensió d i del nombre de *tokens* n . Primer de tot, podem destacar que C3 no varia gaire amb (n, d) i que per a totes les configuracions, $\max_i \|x_i(T)\| \approx 10^{-3}$ indicant que l'estabilitat de l'origen és uniforme.

En canvi els casos C4, C5, C6, C7, C8 presenten un creixement de $\max_i \|x_i\|$ conforme augmenta la dimensió, passant de ~ 1 a baixa dimensió ($d = 3$) a $\sim 4,5$ o $6,5$ a (3200, 64). Els casos indefinits C5 i C7 creixen més de pressa que els altres. A (3200, 64) veiem que C5 té $\max_i \|x_i\| = 6,55$, i C4 té 4,57. També que C7

Cas	(100, 3)	(200, 4)	(300, 5)	(400, 8)	(800, 16)	(1600, 32)	(3200, 64)
C1 (Conservatiu)	26	52	119	112	745	1596	3200
C2 (Dissipatiu sense MLP)	1	1	1	1	1	1	1
C3 (MLP lineal estable)	1	1	1	1	1	1	1
C4 (MLP lineal inestable)	35	90	139	321	800	1600	3200
C5 (MLP lineal indefinit)	25	49	90	179	400	801	1601
C6 (MLP cúbic repulsiu)	8	16	32	209	738	1600	3200
C7 (MLP cúbic indefinit)	9	17	33	141	382	801	1601
C8 (Repulsiu de grau 5)	8	16	32	211	723	1600	3200

Taula 5.5: Nombre d'agrupacions detectades a l'estat final $X(T)$ amb $T = 150$.

$\max_i \|x_i\| = 5,76$, en canvi 4,25 a C6. Això indica que la meitat amb $\alpha_i > 0$ pot allunyar-se més lluny.

La evolució del sistema en els diferents casos a $(n, d) = (100, 3)$ es pot observar molt bé mitjançant un GIF (snapshots a l'annex), il·lustrant aspectes qualitius que les altres gràfiques no transmeten.

El cas C1 s'observa que els *tokens* es concentren en una regió petita al voltant del centre i es reexpandeixen cap a S^2 (rebot). Seguidament formen una estructura allargada, la orientació del sistema va canviant constantment. Aquests esdeveniments es reproduïxen cíclicament.

Els casos lineals C2, C3 i C4 mostren comportaments molt diferents. A C2 (sense MLP) el col·lapse principal es produeix al començar la visualització i el sistema continua contraient-se lentament, a $T = 150$ tots els *tokens* es troben a $\|x_i\| \approx 0,04$. A C3 el comportament és similar però més veloç i concentrat, arribant a $\max_i \|x_i\| \approx 0,001$ s, l'efecte combinat de l'atenció i del MLP estable fa que aquest sigui el cas que més ràpidament col·lapsa al punt fix $X = 0$. A C4, en canvi, els *tokens* es contrauen una mica i es expandeixen continuament de forma lleugera fins a estar distribuïts al voltant de l'esfera unitària, la configuració canvia visiblement durant el primer terç de la simulació.

El cas C5 (lineal indefinit) veiem que un grup de tokens es comporta de forma molt similar que a C3 i l'altra a C4 però rebotant molt més lluny de l'esfera unitària. També estàn de forma més dispersa, les files $\alpha_i < 0$ es contrauen cap a l'origen i les $\alpha_i > 0$ es repel·len. A diferència de C6/C7/C8, C5 no forma agrupacions discretes a baixa dimensió, es detecten grups individuals o de 2 que formen un clúster artificialment. Els *tokens* s'estabilitzen en una distribució on $\|x_i\|$ varia en l'interval aproximat $[0, 1,5]$. En altes dimensions, els dos grups es diferencien de forma molt clara a l'histograma de la Figura 5.11, amb una part a prop de l'origen i l'altra de $\|x_i\| \approx 6,5$.

Els casos no lineals C6 i C8 produeixen agrupacions discretes dels *tokens* al voltant de S^2 . A C6 els *tokens* es contrauen i reboten sense estructura visible per $t < 6$, entre $t \approx 6$ i $t \approx 10$ comencen a aparèixer agrupacions, per $t \approx 15$ –20 les agrupacions són clarament visibles amb exactament 8 grups. Per $t \geq 25$ la configuració es manté estable. A C8 l'aparició de les agrupacions és més lenta (entorn de $t \approx 15$ –20), tots els tokens estàn visiblement agrupats per $t \approx 130$.

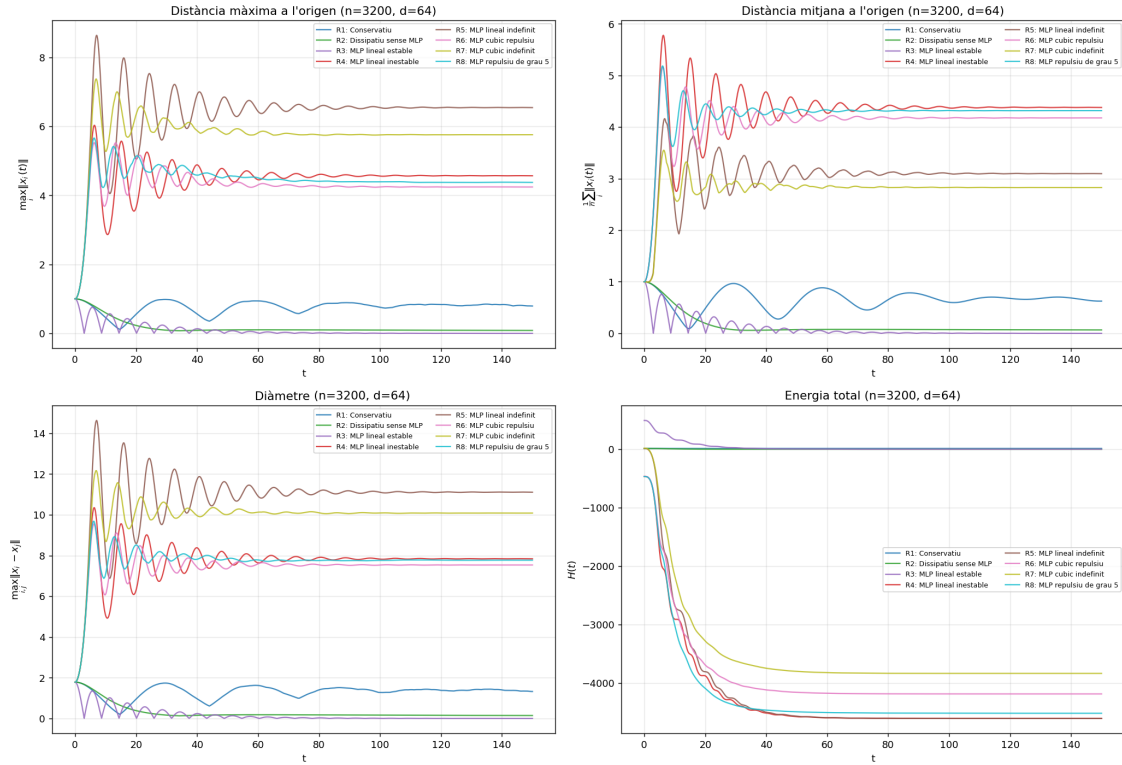


Figura 5.9: Gràfiques per als vuit casos, amb $(n, d) = (3200, 64)$. **Dalt esquer-
ra:** $\max_i \|x_i(t)\|$. **Dalt dreta:** Distància mitjana a l'origen $\frac{1}{n} \sum_i \|x_i(t)\|$. **Baix
esquerra:** Diàmetre del sistema $\max_{i,j} \|x_i - x_j\|$. **Baix dreta:** energia total $H(t)$.

El cas **C7** (cúbic indefinit) mostra una variant del comportament de C6, en lloc d'agrupacions de mida similar, s'observa una agrupació major prop del centre (associada a les files amb $\alpha_i < 0$) i diverses agrupacions com les de C6 però, més petites (associades a les files amb $\alpha_i > 0$). Hi ha 9 grups, un central ($\alpha_i < 0$) i 8 ($\alpha_i > 0$), corresponents al mateix patró que veiem a C6.

La taula 5.5 consisteix en el nombre d'agrupacions a la configuració final utilitzant clustering jeràrquic. Els valors confirmen els patrons que hem descrit i revelen alguns de nous.

Els casos C2 i C3 són els únics on es detecten 1 agrupació a totes les configuracions, corresponent al col·lapse al voltant de l'origen. En canvi, cap altre cas manté un nombre constant de grups mentre augmentem les n o d , a C4, C5, C6, C7 i C8 el nombre d'agrupacions va creixent des de $\sim 8 - 35$ a $(100, 3)$ fins a $\sim 1600 - 3200$ a $(3200, 64)$. A C1 (conservatiu) el nombre d'agrupacions detectades també creix amb la dimensió, però això no s'ha d'interpretar com a formació de *clusters*, ja que no hi ha convergència. Per tant, no té sentit fixar-nos en C1 per aquest apartat. Cal destacar que aquestes agrupacions poden ser només d'un token.

A C5 i C7 (els dos indefinits), els seus nombres d'agrupacions són aproximadament la meitat dels de C4 i C6 respectivament. Això és coherent amb que en els casos indefinits només la meitat dels *tokens* ($\alpha_i > 0$) s'allunya de l'origen, mentre que l'altra meitat col·lapsa en aquest.

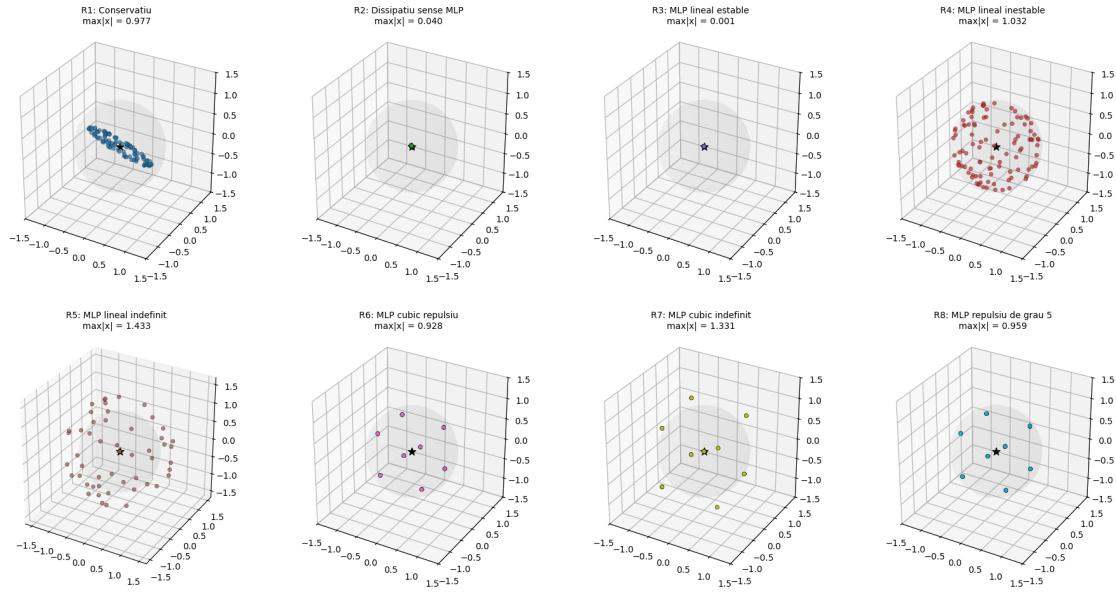


Figura 5.10: Estats finals dels vuit casos a $T = 150$ amb $n = 100$ *tokens* en $\mathbb{R}^3$.

A altes dimensions ens trobem una dificultat que era predictable, detectar ~ 3200 agrupacions a $(3200, 64)$, tantes com *tokens*, no significa que el sistema hagi perdut l'estructura, sinó que a alta dimensió els *tokens* es separen tant entre si que cadascun forma el seu propi *cluster*, això passa perquè la densitat de tokens és sigui tan petita que la distribució d'aquests i les trajectòries siguin tant diferents que l'estructura no sigui visible.

A baixa dimensió veiem els casos no lineals (C6, C7, C8) mostren un patró interessant. Observem la relació $\#clusters \approx 2^d$ per C6 i C8, i $2^d + 1$ al cas indefinit C7 (un grup central addicional). Seria interessant estudiar com el MLP i l'Attention generen aquesta relació.

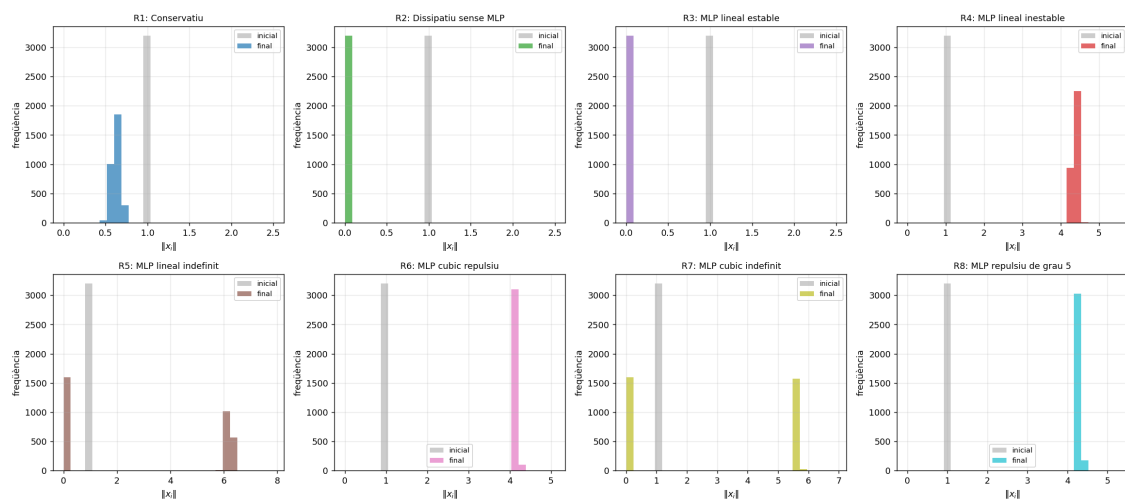


Figura 5.11: Histogrames de la distància a l'origen $\|x_i\|$ per als vuit casos amb $(n, d) = (3200, 64)$ (gris: distribució inicial sobre S^{d-1} , color: distribució final).

Capítol 6

Conclusions

En aquest treball hem estudiat l'arquitectura del *Transformer* des d'una perspectiva matemàtica, interpretant-lo com un sistema de partícules. Hem començat per l'estructura del bloc, trobant l'EDO de primer ordre que el modela i deduint la versió de segon ordre per controlar el creixement de les trajectòries. Aquesta l'hem aconseguida a partir de la variant del mètode Symplectic-Euler de [4] amb una reparametrització que introdueix Δt . En fer el pas a la dinàmica contínua, en perdre's el terme de *lookahead*, la versió contínua de l'esquema coincideix amb el *heavy-ball* de Polyak.

Aquest segon model ha estat el nucli del treball i l'hem analitzat teòricament, demostrant que les trajectòries estan acotades i classificant els diferents punts d'equilibri segons la hessiana del potencial, el tractament del cas degenerat (hessiana singular) queda com a l'ínia natural de continuació. També hem investigat comportaments mitjançant les simulacions numèriques que han confirmat els diferents resultats teòrics i hem pogut veure quines poden ser possibles línies futures per expandir el treball.

Al llarg del document hem notat que la construcció de la funció d'energia total H implica restriccions fortes sobre els elements clau del *Transformer*, l'*Attention* i el MLP. Hem pogut observar les diferents formes de *clustering* segons el MLP, lineal o no, mostrades per les simulacions. Queda com a extensió del treball l'estudi profund de per què es formen aquests *clusters* i diferents patrons com l'últim de 2^d agrupacions que hem vist. Es podria comparar amb diferents tipus d'*Attention* sempre que es trobin més que puguin complir les suposicions del treball. També, amb més recursos computacionals podem indagar millor en el comportament de la dinàmica amb n tokens i d molt més propers a la realitat.

Annex A

En aquest Annex mostrem els diferents snapshots (cada 3 segons) per veure la evolució del sistema en els diferents casos a $(n, d) = (100, 3)$ del experiment 5.

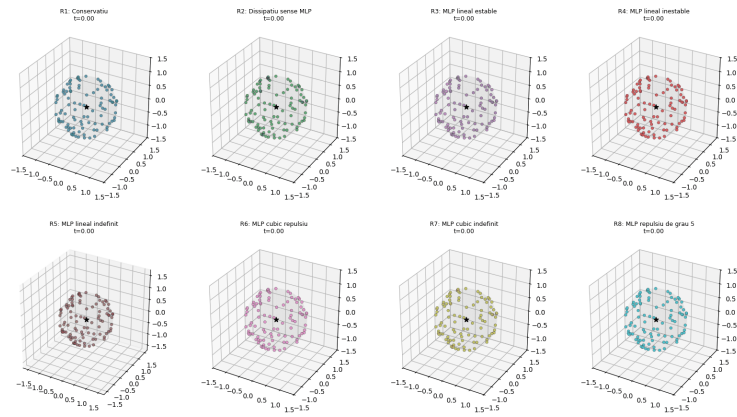


Figura A.1: Snapshot 1

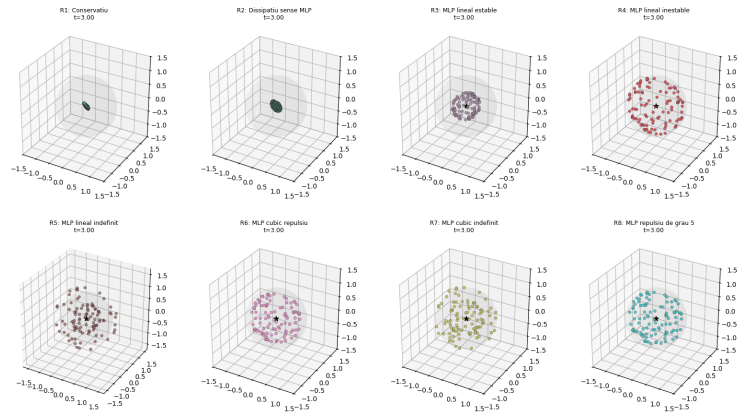


Figura A.2: Snapshot 2

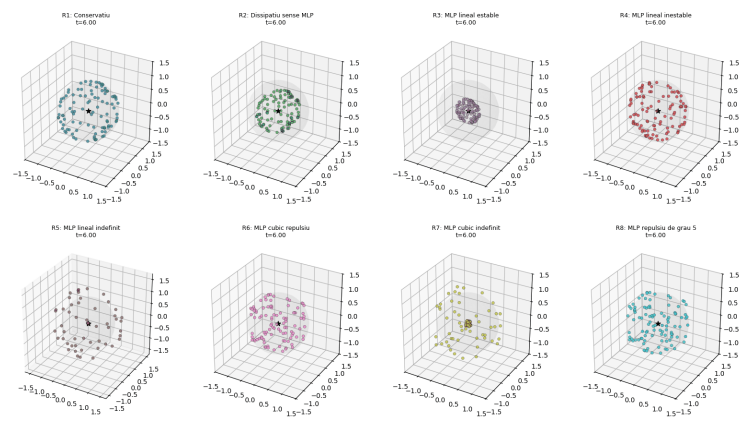


Figura A.3: Snapshot 3

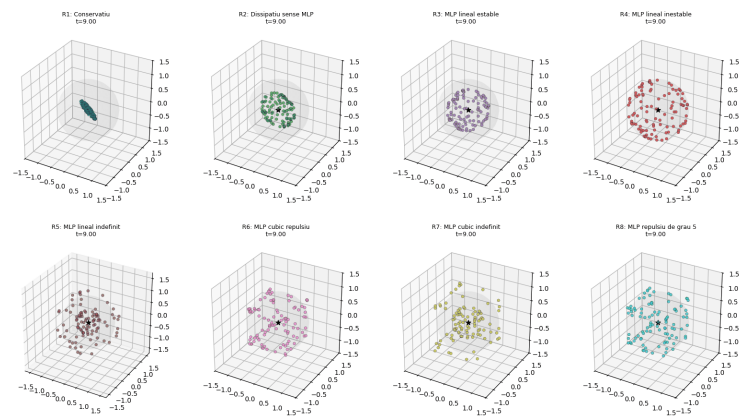


Figura A.4: Snapshot 4

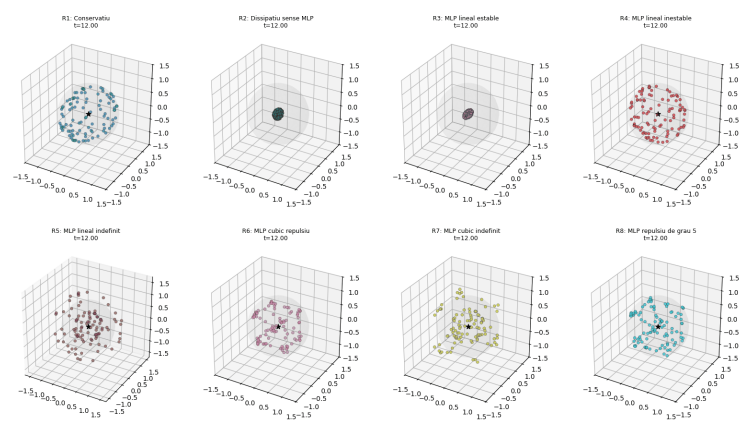


Figura A.5: Snapshot 5

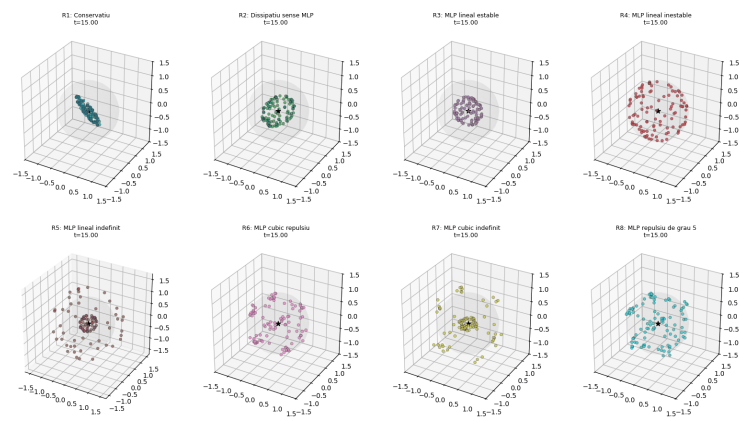


Figura A.6: Snapshot 6

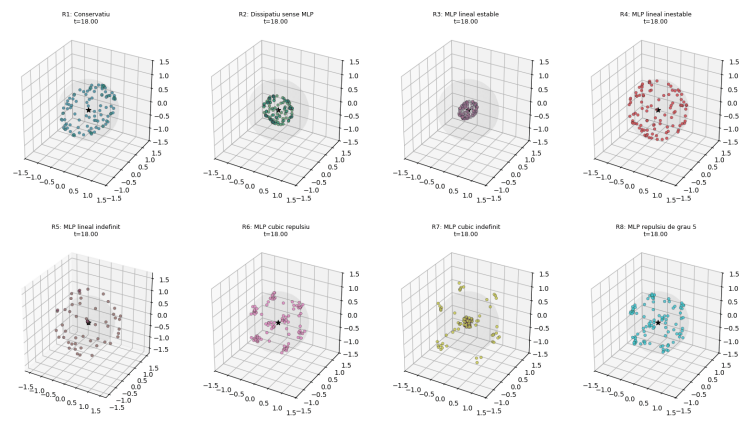


Figura A.7: Snapshot 7

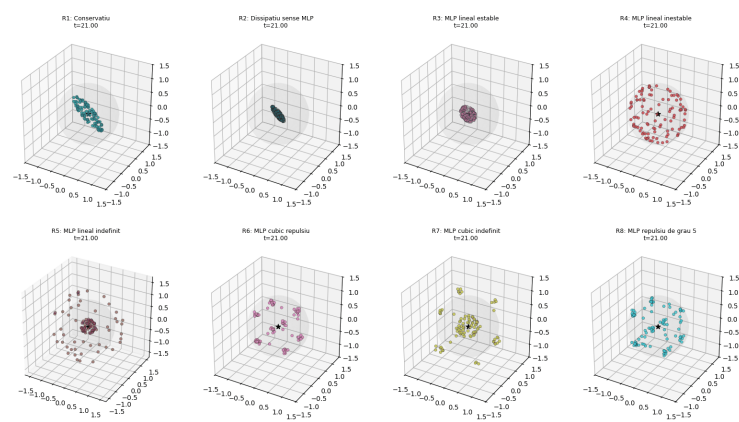


Figura A.8: Snapshot 8

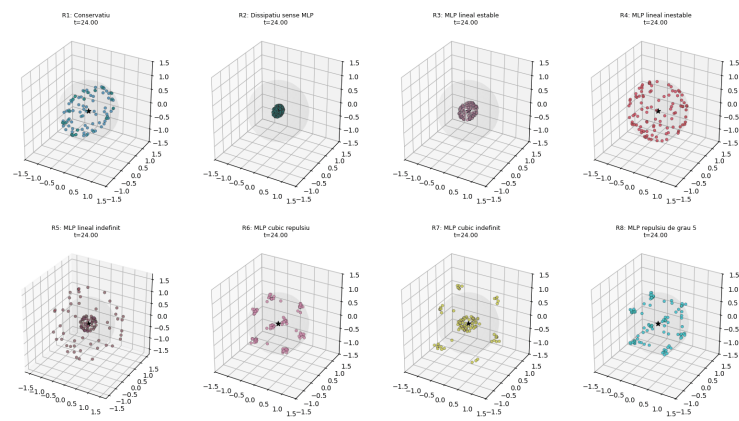


Figura A.9: Snapshot 9

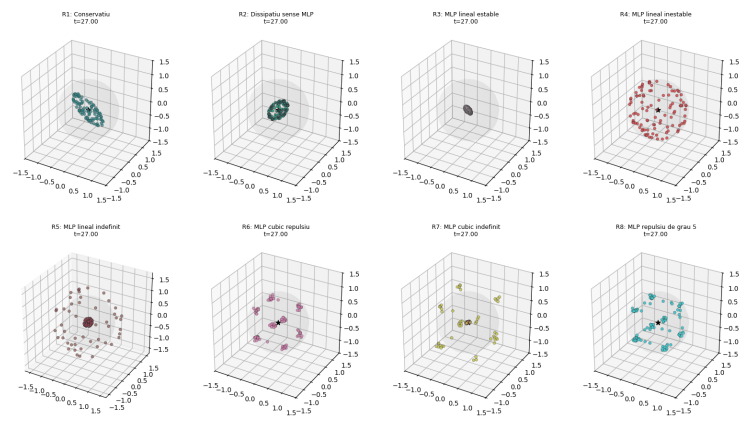


Figura A.10: Snapshot 10

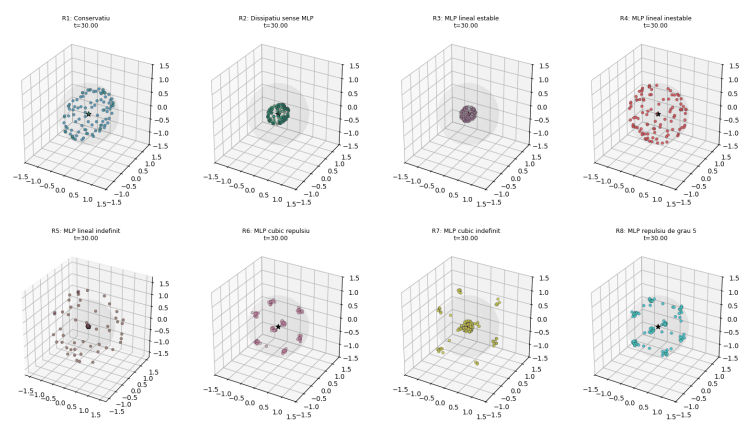


Figura A.11: Snapshot 11

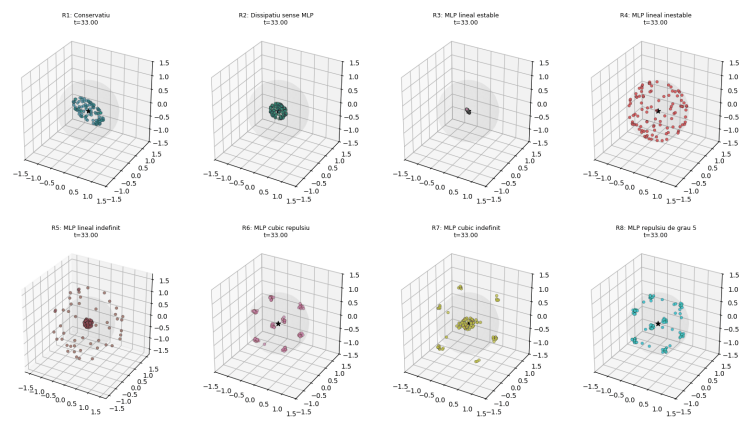


Figura A.12: Snapshot 12

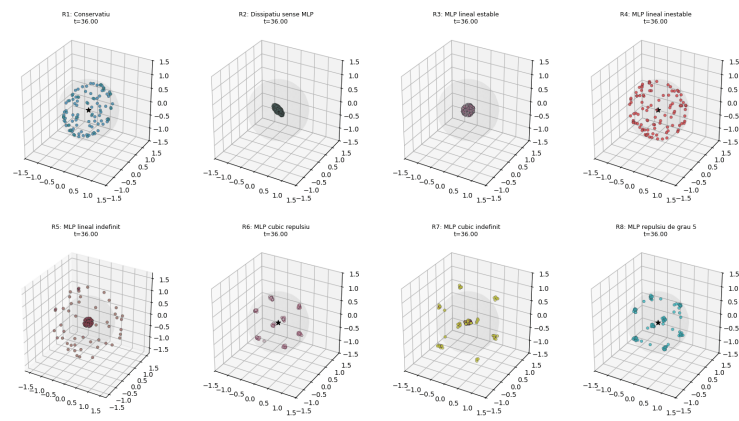


Figura A.13: Snapshot 13

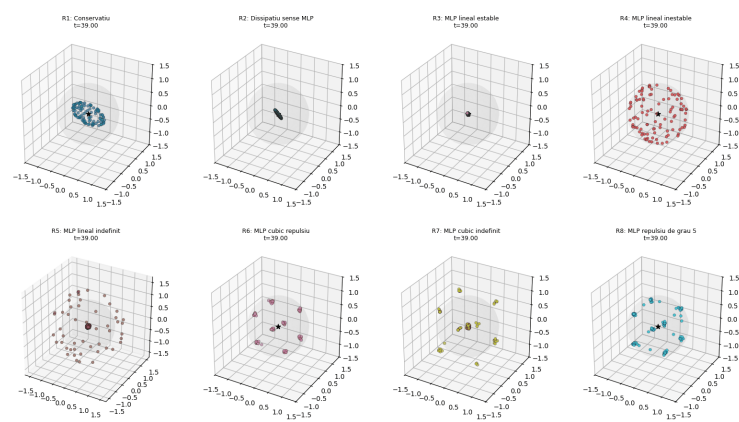


Figura A.14: Snapshot 14

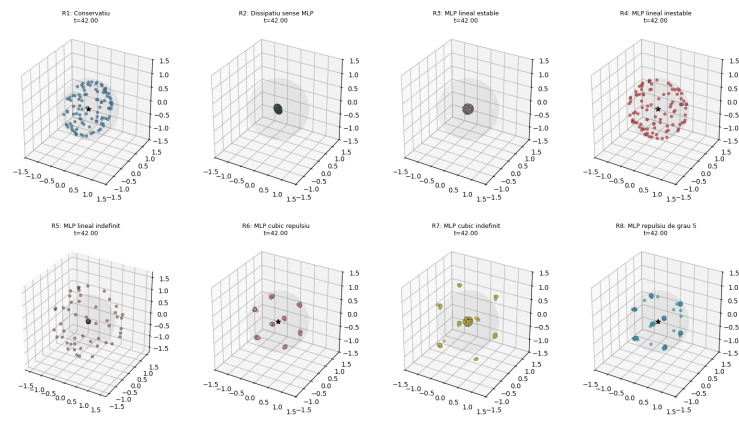


Figura A.15: Snapshot 15

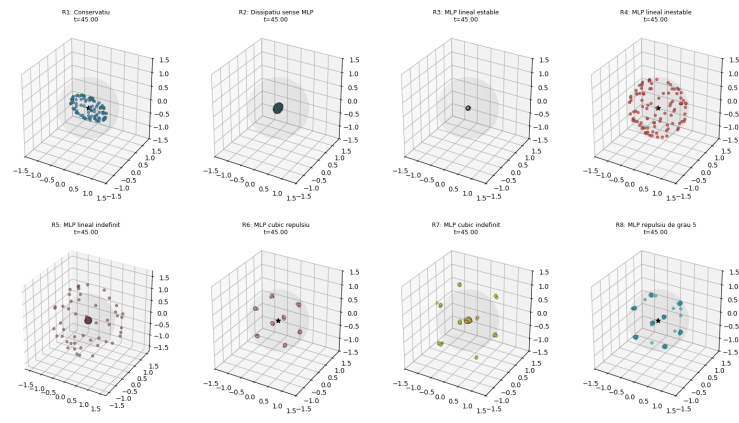


Figura A.16: Snapshot 16

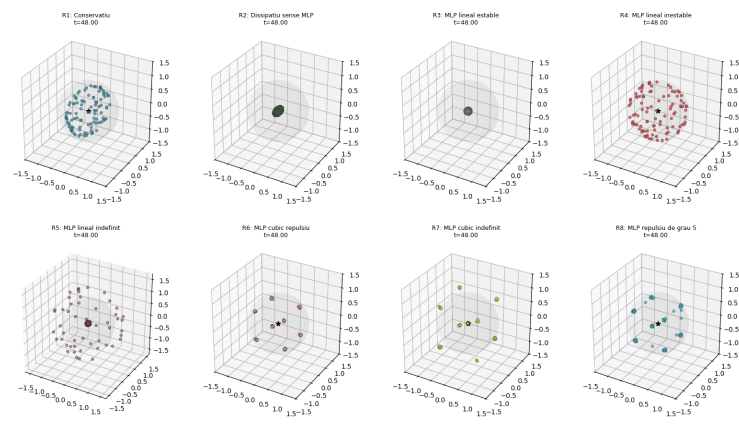


Figura A.17: Snapshot 17

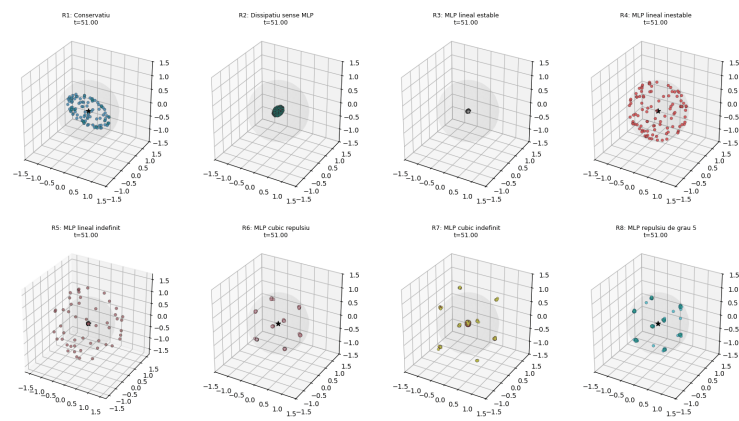


Figura A.18: Snapshot 18

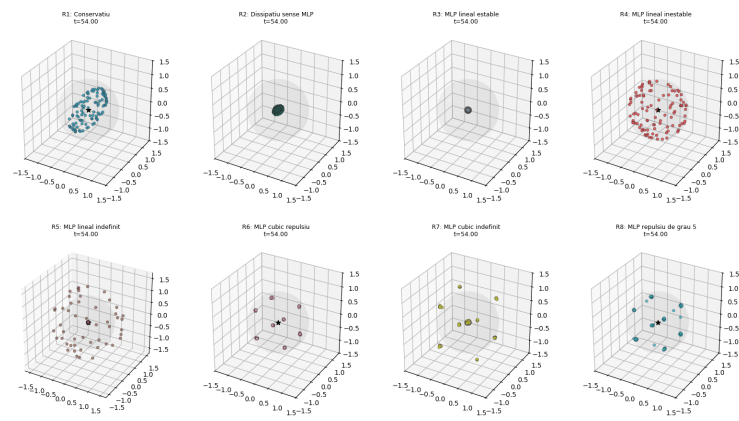


Figura A.19: Snapshot 19

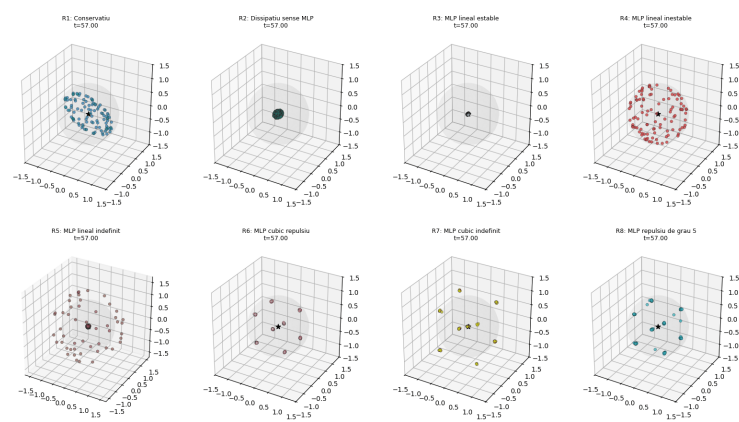


Figura A.20: Snapshot 20

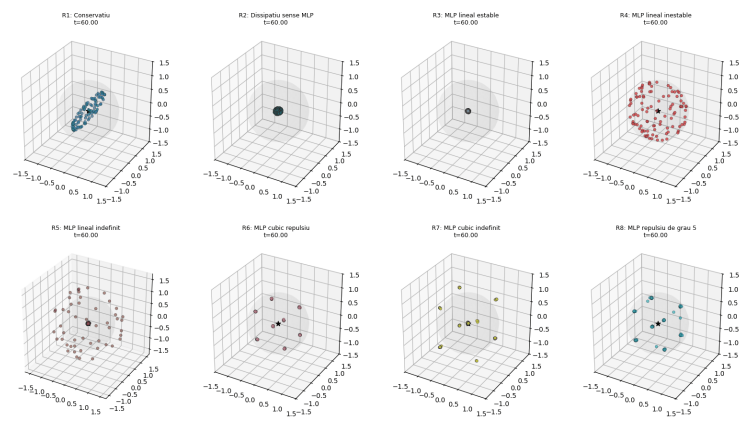


Figura A.21: Snapshot 21

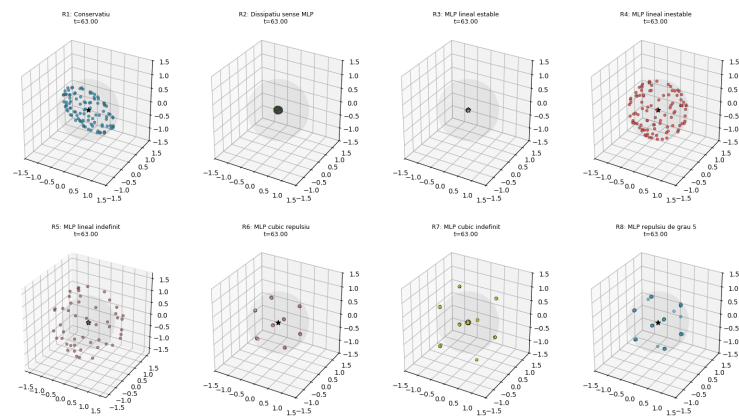


Figura A.22: Snapshot 22

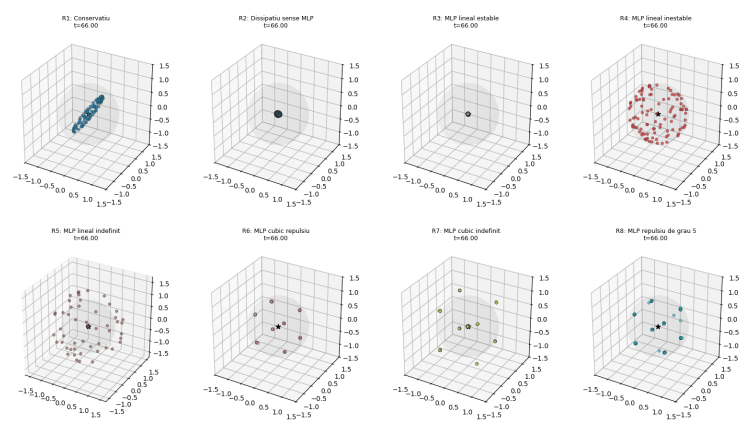


Figura A.23: Snapshot 23

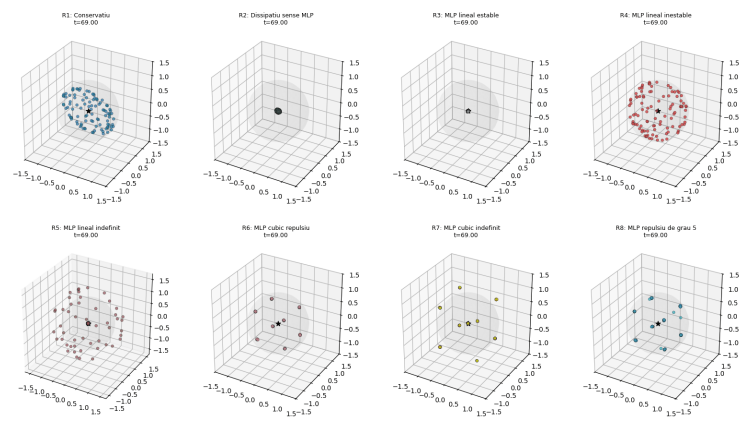


Figura A.24: Snapshot 24

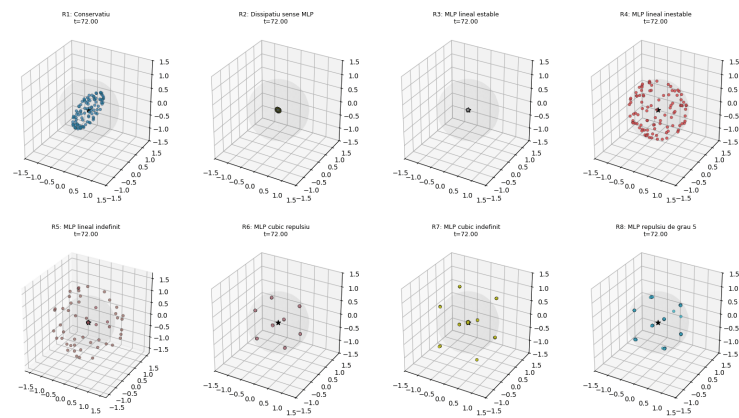


Figura A.25: Snapshot 25

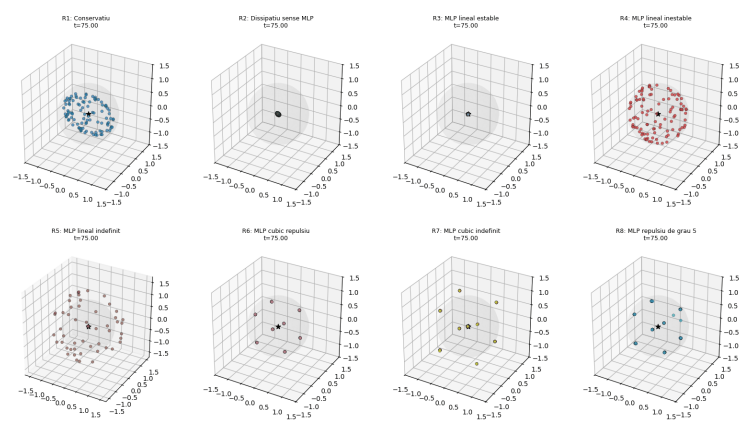


Figura A.26: Snapshot 26

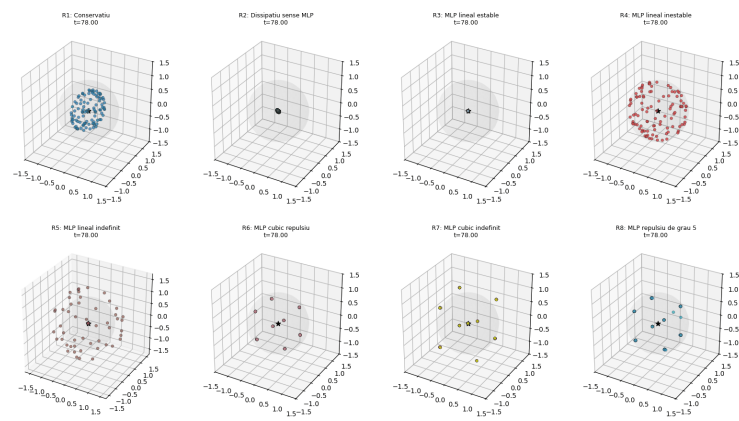


Figura A.27: Snapshot 27

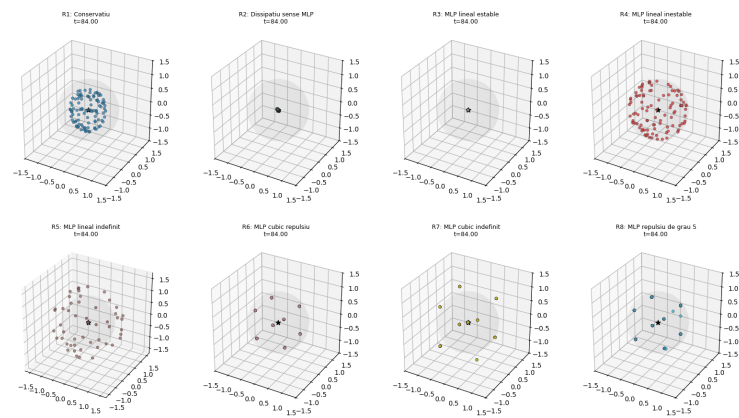


Figura A.28: Snapshot 28

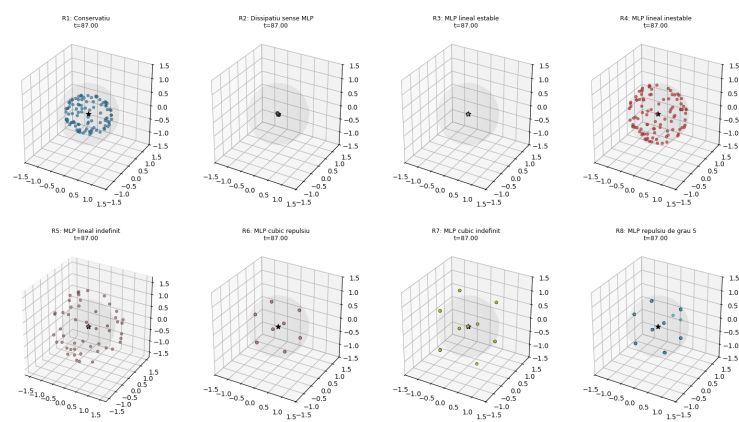


Figura A.29: Snapshot 29

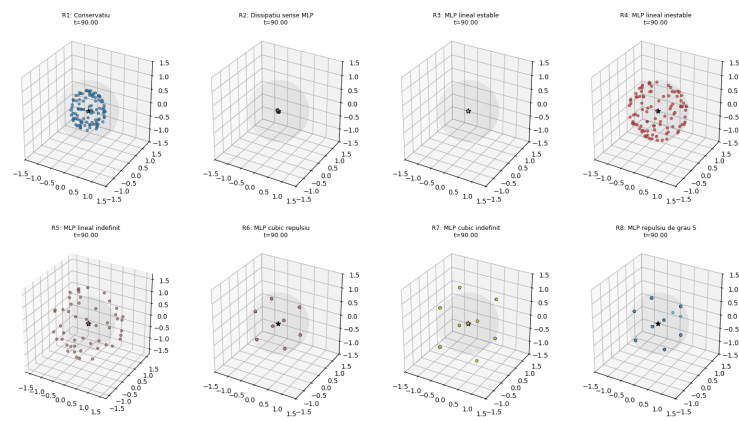


Figura A.30: Snapshot 30

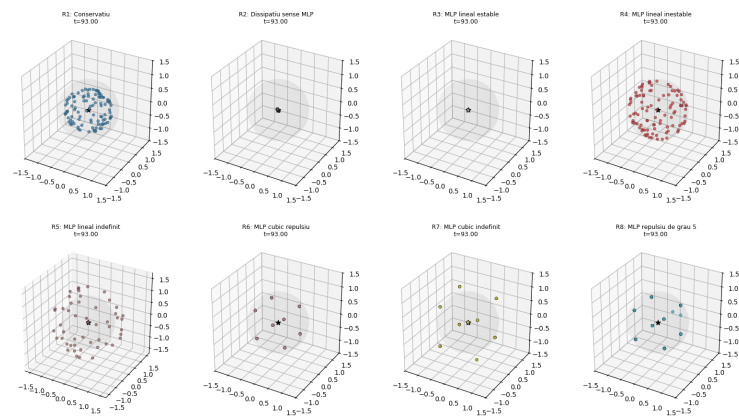


Figura A.31: Snapshot 31

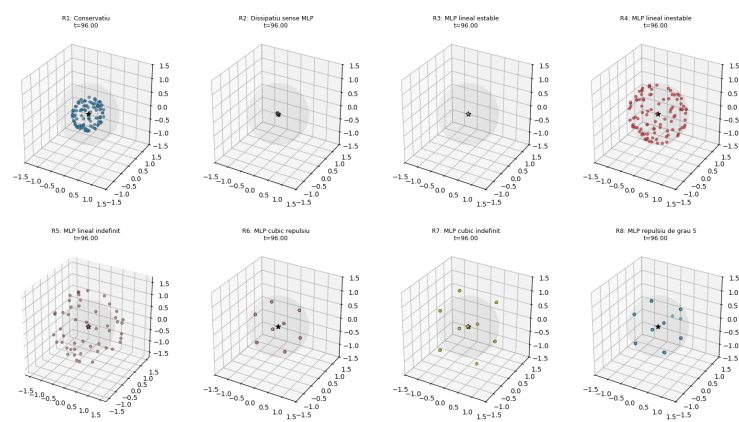


Figura A.32: Snapshot 32

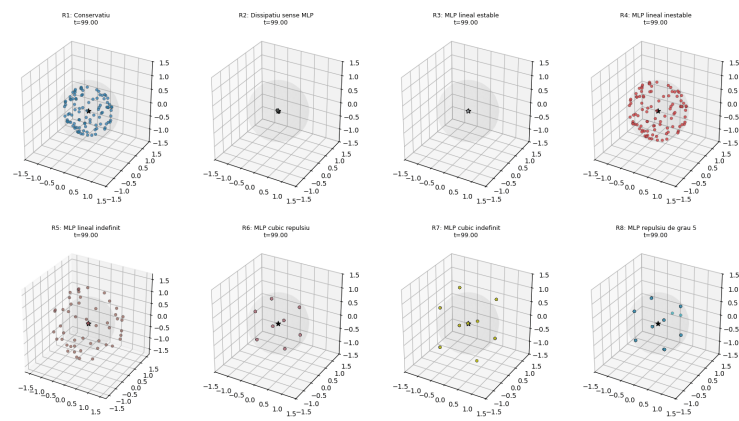


Figura A.33: Snapshot 33

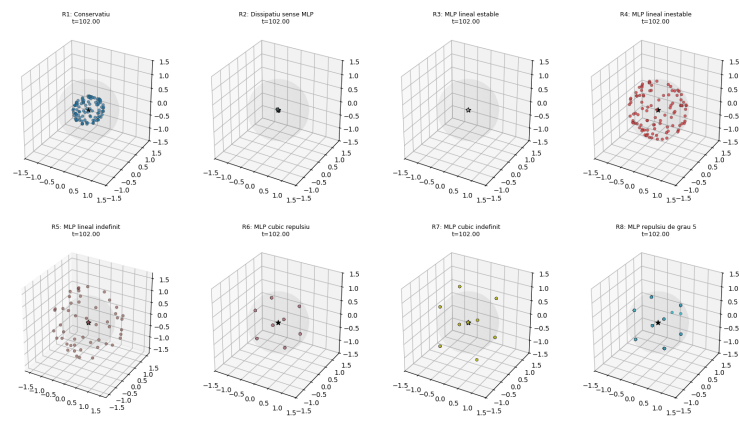


Figura A.34: Snapshot 34

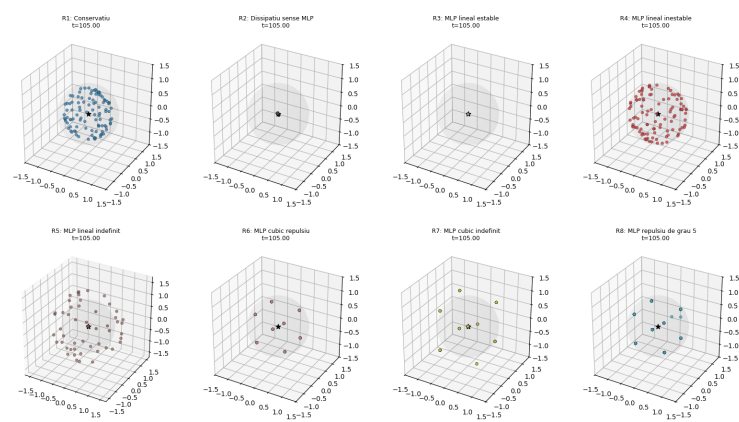


Figura A.35: Snapshot 35

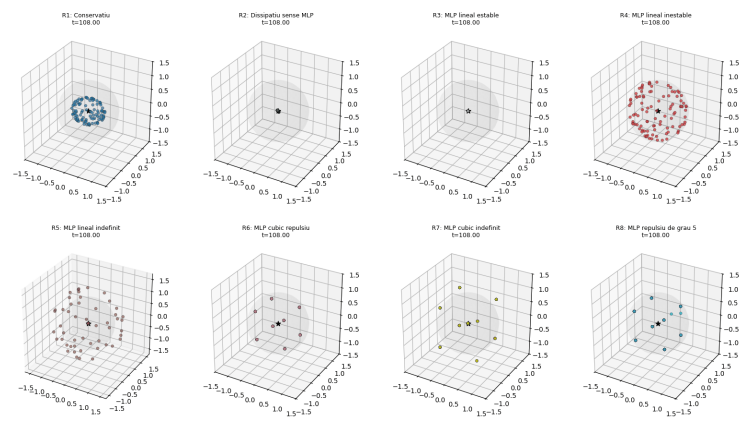


Figura A.36: Snapshot 36

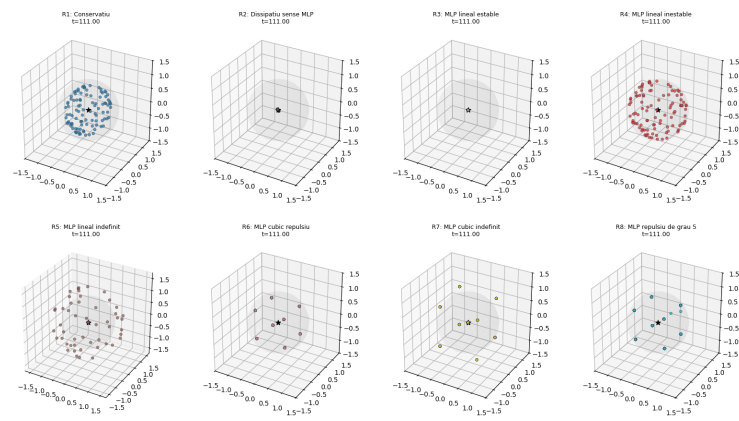


Figura A.37: Snapshot 37

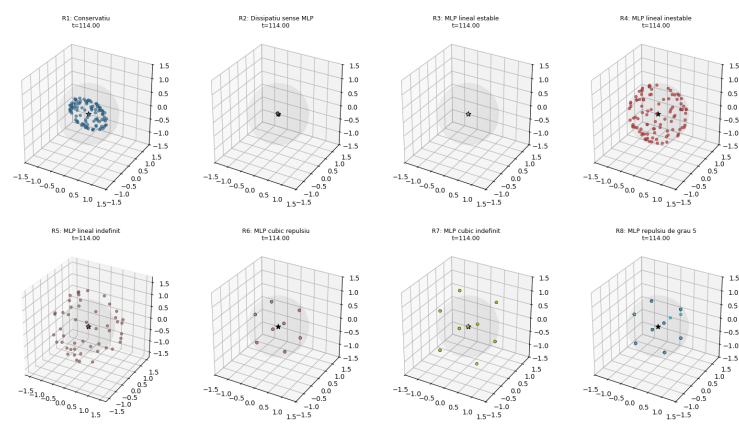


Figura A.38: Snapshot 38

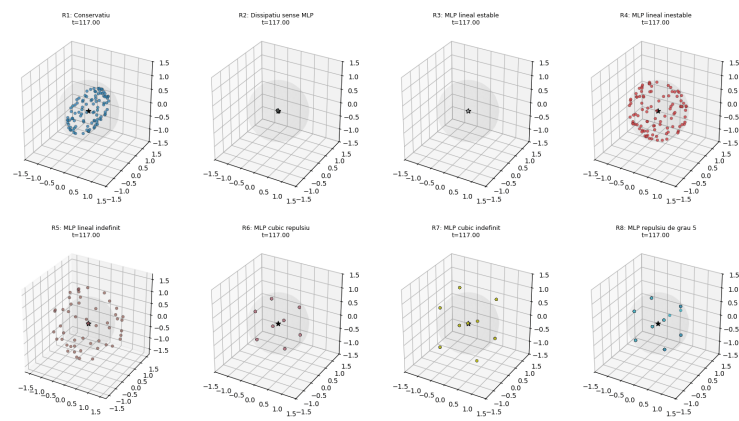


Figura A.39: Snapshot 39

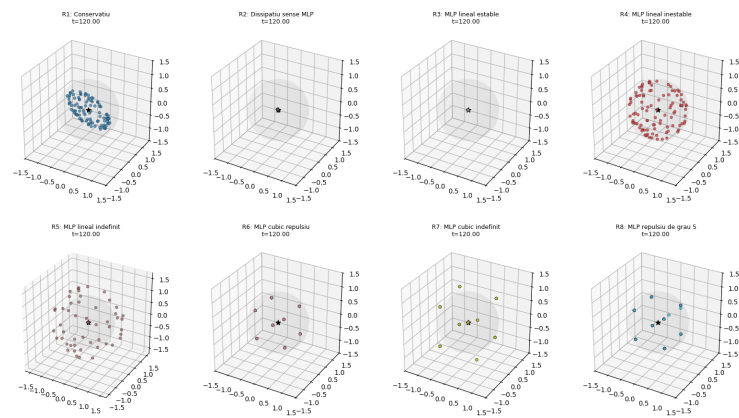


Figura A.40: Snapshot 40

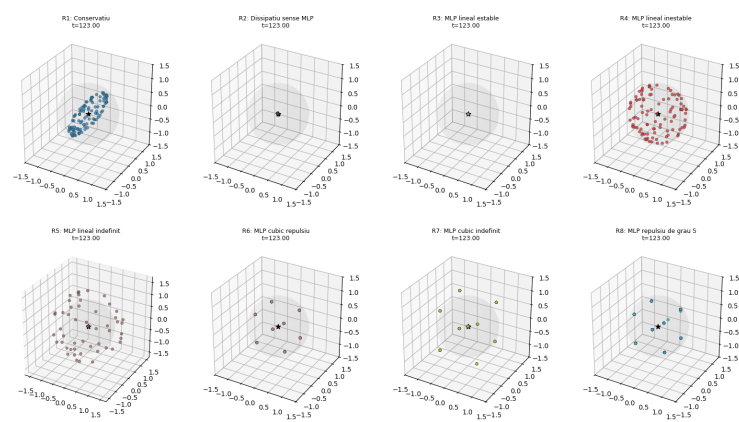


Figura A.41: Snapshot 41

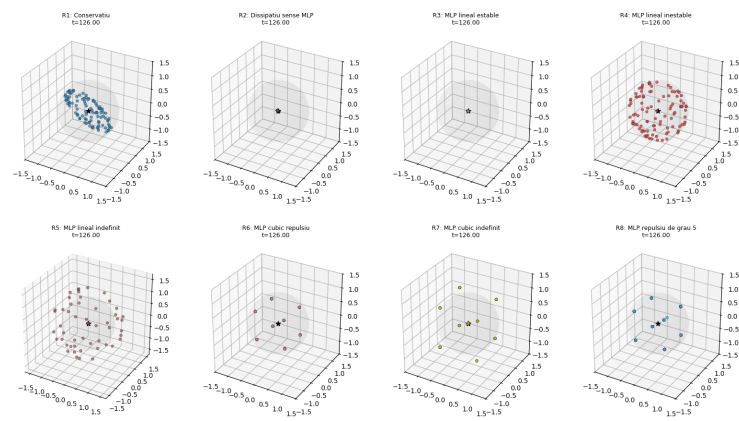


Figura A.42: Snapshot 42

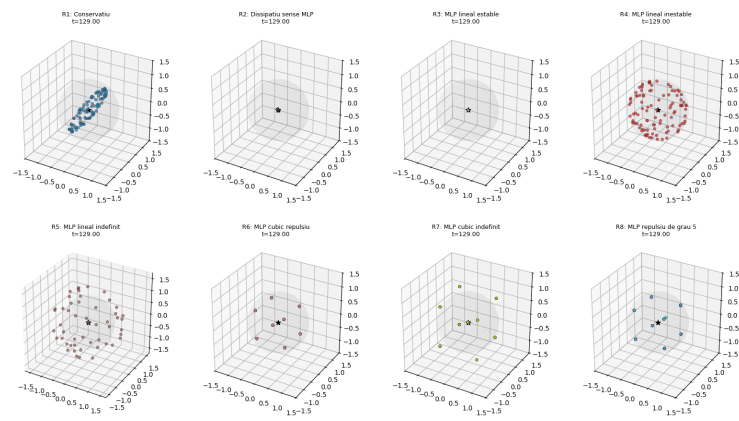


Figura A.43: Snapshot 43

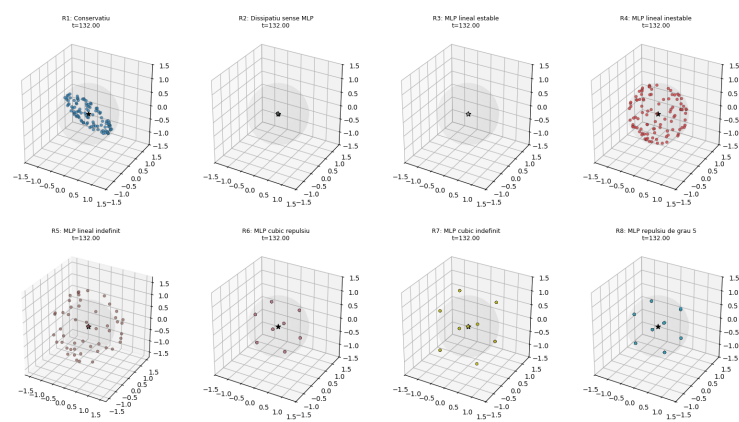


Figura A.44: Snapshot 44

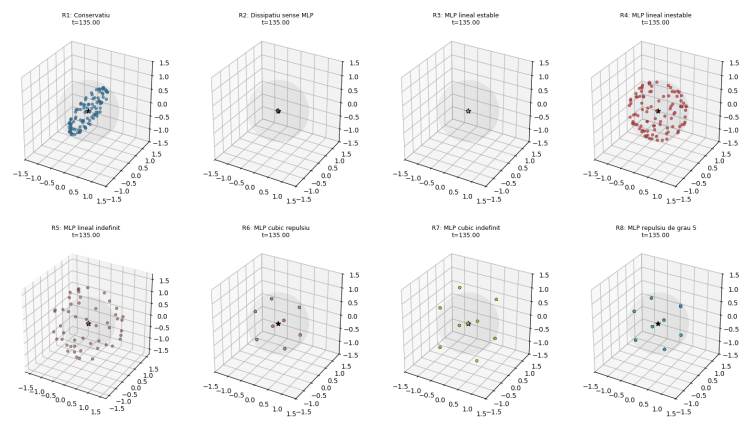


Figura A.45: Snapshot 45

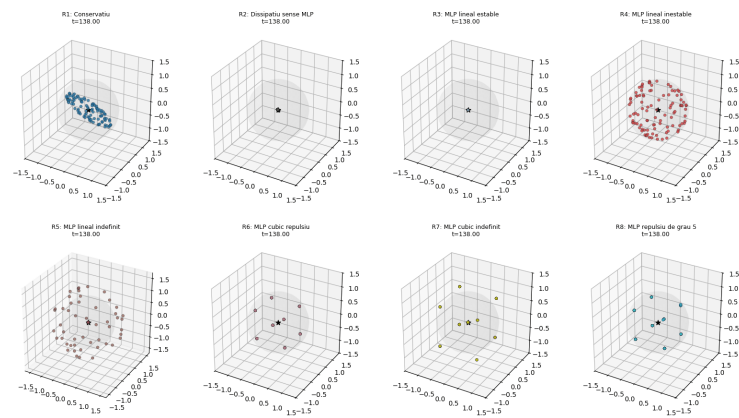


Figura A.46: Snapshot 46

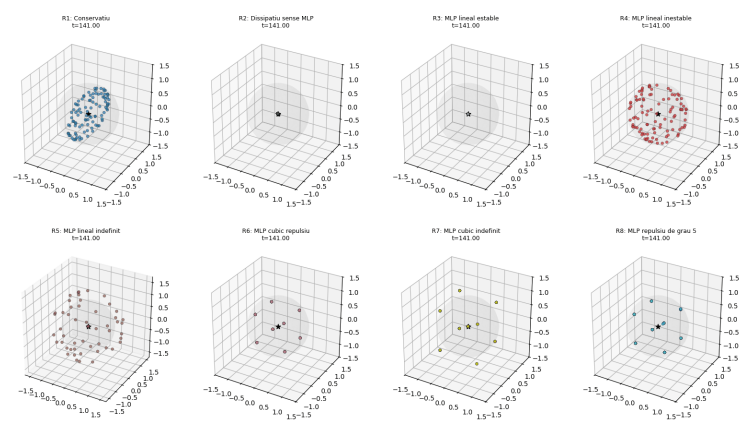


Figura A.47: Snapshot 47

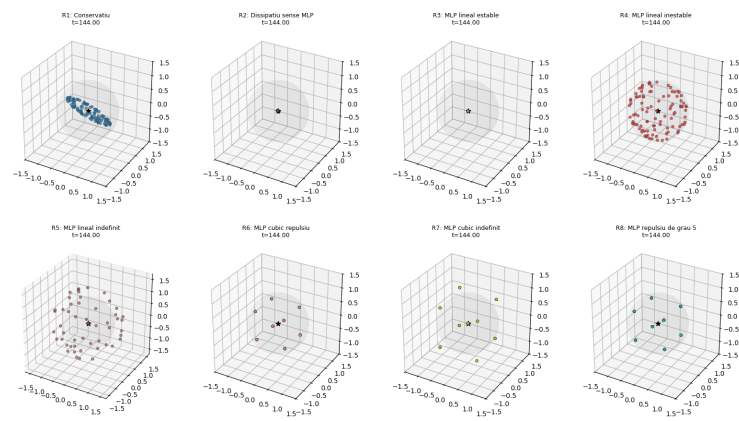


Figura A.48: Snapshot 48

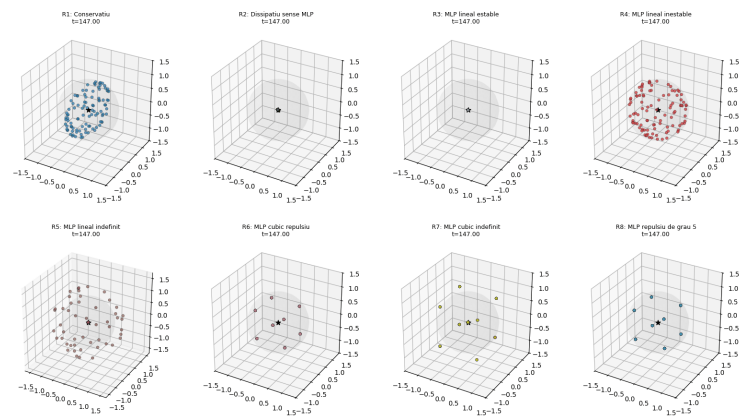


Figura A.49: Snapshot 49

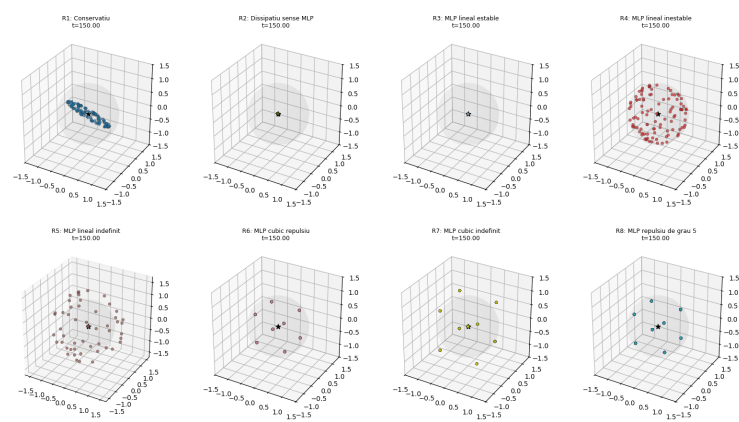


Figura A.50: Snapshot 50

Bibliografia

- [1] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I.: *Attention Is All You Need*, arxiv.org/abs/1706.03762, 2017.
- [2] Geshkovski, B.; Letrouit, C.; Polyanskiy, Y.; Rigollet, P.: *A mathematical perspective on Transformers*, <https://arxiv.org/abs/2312.10794>, 2023.
- [3] Geshkovski, B.; Rigollet, P.; Ruiz-Balet, D.: *Measure-to-measure interpolation using Transformers*, arxiv.org/abs/2411.04551, 2024.
- [4] Zimin, A.; Polyanskiy, Y.; Rigollet, P.: *YuriiFormer: A Suite of Nesterov-Accelerated Transformers*, arxiv.org/abs/2601.23236, 2026.
- [5] Polyak, B. T.: *Some methods of speeding up the convergence of iteration methods*, *USSR Computational Mathematics and Mathematical Physics*, papers.baulab.info/papers/also/Polyak-1964.pdf, 1964.
- [6] Teschl, G.: *Ordinary Differential Equations and Dynamical Systems*, Graduate Studies in Mathematics, vol. 140, American Mathematical Society, Providence, Rhode Island, 2012.
- [7] Khalil, H. K.: *Nonlinear Systems*, 3a edició, Prentice Hall, Upper Saddle River, New Jersey, 2002.
- [8] Hairer, E., Lubich, C., Wanner, G.: *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer 2nd edition, 2006.
- [9] Katharopoulos, A.; Vyas, A.; Pappas, N.; Fleuret, F.: *Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention*, arxiv.org/abs/2006.16236, 2020.
- [10] Álvarez-López, A.; Geshkovski, B.; Ruiz-Balet, D.: *Perceptrons and localization of attention's mean-field landscape*, <https://arxiv.org/abs/2601.21366>, 2026.

- [11] Geshkovski, B.; Letrouit, C.; Polyanskiy, Y.; Rigollet, P.: *The emergence of clusters in self-attention dynamics*, arxiv.org/abs/2305.05465, 2023.
- [12] Pinkus, A.: *Approximation theory of the MLP model in neural networks*, *Acta Numerica*, pinkus.net.technion.ac.il/files/2021/02/acta.pdf, 1999.
- [13] Clop, A.: Càlcul diferencial en diverses variables, Apunts de classe, Universitat de Barcelona, 2023.
- [14] Gonchenko, M.: Mètodes numèrics II, Apunts de classe, Universitat de Barcelona, 2025.