

Ensemble Controllability of Control-Linear Mean-Field Transformers

Workshop **Machine Learning + Simulation**

Jean-Baptiste Caillau¹ Chih-Kang Huang^{1, 2} Ludovic Rifford¹

¹Université Côte d'Azur, CNRS, Inria, LJAD

²Université Côte d'Azur, Inria (MCTAO), CNRS, LJAD

October 13th 2026



A ResNet with L layers is defined as follows. Fix $T > 0$ and set $\text{dt} = \frac{T}{L}$. For $l = 1, \dots, L$,

$$\begin{cases} x^{(l)} = x^{(l-1)} + \text{dt} \sigma(W^{(l)} x^{(l-1)} + B^{(l)}), \\ x^{(0)} = x^0 \in \mathbb{R}^d. \end{cases}$$

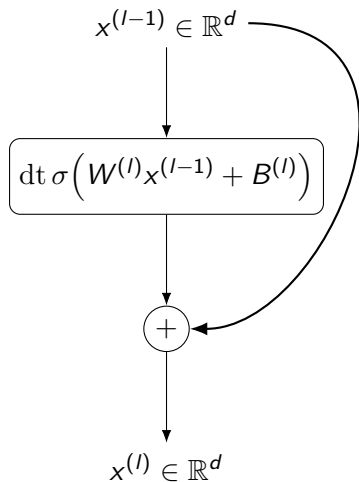
The parameters of layer l are

$$\theta^{(l)} = \{W^{(l)} \in \mathbb{R}^{d \times d}, B^{(l)} \in \mathbb{R}^d\},$$

and

$$\sigma : \mathbb{R}^d \rightarrow \mathbb{R}^d$$

is an *activation function* (ReLU, tanh, ...).



From ResNets to Neural ODEs

(E et al. '17, Chen et al. '18) A ResNet can be viewed as a forward Euler discretization of a *Neural ODE*

As $L \rightarrow \infty$ (and $dt = T/L \rightarrow 0$), $x^{(l)} = x^{(l-1)} + dt \sigma(W^{(l)}x^{(l-1)} + B^{(l)})$ formally converges to

$$\boxed{\begin{cases} \dot{x}(t) = \sigma(W(t)x(t) + B(t)) & t \in [0, T], \\ x(0) = x^0. \end{cases}} \quad (1)$$

The time-dependent parameters

$$\theta(t) = (W(t), B(t))$$

play the role of the *control* and are typically taken to be piecewise constant.

The flow map induced by the system is a diffeomorphism isotopic to the identity

$$\text{Diff}_0(\mathbb{R}^d) \ni \Phi_T^\theta : x^0 \longmapsto x(T).$$

Network expressivity $\longleftrightarrow$ *Ensemble controllability of the Neural ODE.*

Ensemble Controllability

Finite exact controllability. Given two N -tuples of distinct points $(x^i)_{i=1}^N$ and $(y^i)_{i=1}^N$, can we choose θ such that

$$\Phi_T^\theta(x^i) = y^i, \quad i = 1, \dots, N?$$

Approximate controllability. Given a target map $\Phi \in \text{Diff}_0(\mathbb{R}^d)$, for $\epsilon > 0$ and every compact $K \subset \mathbb{R}^d$, can we choose a piecewise-constant control θ such that

$$\|\Phi_T^\theta - \Phi\|_{C^0(K)} \leq \epsilon?$$

Controllability results for Neural ODEs (1): Tabuada & Gharesifard (2020), and Zuazua et al. (2022– for $\sigma = \text{ReLU}$).

What about other neural network architectures?

Transformer Dynamics

Following Geshkovski *et al.* (2023), consider N tokens

$$X_N(t) = (x_1(t), \dots, x_N(t)) \in (\mathbb{S}^{d-1})^N, \quad d \geq 3,$$

evolving according to

$$\dot{x}_i = P_{x_i}^\perp \left(\sum_{j=1}^N \alpha_{Q(t)}^\beta(x_i, x_j) V(t) x_j \right) \quad (2)$$

with $Q(t), V(t) \in \mathbb{R}^{d \times d}$ (Query-Key, Value) and the *self-attention mechanism*

$$\alpha_Q^\beta(x_i, x_j) = \frac{e^{\beta \langle x_i, Q x_j \rangle}}{\sum_{\ell=1}^N e^{\beta \langle x_i, Q x_\ell \rangle}}.$$

with $\beta \geq 0$ the *inverse temperature* and $P_x^\perp(y) = y - \langle x, y \rangle x \in T_x \mathbb{S}^{d-1}$ the *layer normalization*.

Why Couple Attention with ResNet?

Existing controllability results for Transformers (e.g. Abu et al. '22, Zuazua et al. '26) typically consider attention dynamics (2) *coupled with ResNet/Neural ODE* (1).

Geshkovski–Rigollet–Ruiz-Balet '26

Measure-to-measure controllability/interpolation.

Self-Attention: separate the supports of different input measures $\mu_0^1, \dots, \mu_0^N$:

$$\text{supp } \mu_{\text{sep}}^i \cap \text{supp } \mu_{\text{sep}}^j = \emptyset, \quad i \neq j.$$

ResNet: exploit this spatial separation to steer the different regions independently.

Our motivation

Can *attention alone* achieve finite-particle ensemble controllability?

Transformer as a Control-Linear System

Goal: cast the Transformer dynamics into a form amenable to geometric control theory.

Fix $Q = \text{Id}_d$ and choose

$$V(t) = \sum_{k=1}^m u_k(t) b_k b_k^\top, \quad b_k \in \mathbb{R}^d.$$

Then the dynamics becomes

$$\dot{x}_i = \sum_{k=1}^m u_k(t) \underbrace{\langle b_k, M_i^\beta(X) \rangle}_{G_{k,i}(X)} P_{x_i}^\perp(b_k), \quad \text{with } M_i^\beta(X) = \sum_{j=1}^N \alpha^\beta(x_i, x_j) x_j.$$

Hence, for $X = (x_1, \dots, x_N) \in (\mathbb{S}^{d-1})^N$,

$$\dot{X}(t) = \sum_{k=1}^m u_k(t) G_k(X(t)).$$

Transformer parameters $u_k(t)$ become *controls* $\implies$ Lie-algebraic controllability tools apply.

(Single-Particle) Control-Linear Systems

Consider on a manifold M

$$\dot{x}(t) = \sum_{k=1}^m u_k(t) g_k(x(t)), \quad g_k \in \text{Vect}(M).$$

For vector fields f, g , the *Lie bracket* $[f, g](x) = Dg(x)f(x) - Df(x)g(x)$, and $\text{Lie}\{g_1, \dots, g_m\}$ = vector fields generated by $g_1, \dots, g_m$ and their iterated Lie brackets.

Chow–Rashevskii Theorem

If the *Hörmander condition* holds,

$$\text{Lie}\{g_1, \dots, g_m\}(x) = T_x M, \quad \forall x \in M,$$

then the system is controllable on every connected component of M .

Thus, any x_0 can be steered to any x_f using piecewise-constant controls.

Is Hörmander's condition sufficient for ensemble controllability?

An Obstruction: Antipodal Points

Consider $N = 2$ with

$$x_1(0) = -x_2(0) = x \in \mathbb{S}^{d-1}.$$

For $X = (x, -x)$,

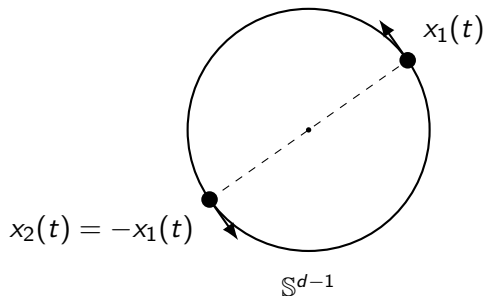
$$M_2^\beta(X) = -M_1^\beta(X), \quad P_{-x}^\perp(b_k) = P_x^\perp(b_k).$$

Hence

$$\dot{x}_2(t) = -\dot{x}_1(t), \quad \boxed{x_2(t) = -x_1(t) \quad \forall t.}$$

Thus, the antipodal relation is invariant and arbitrary pairs cannot be reached.

We do not have ensemble controllability even when the underlying single-particle system satisfies Hörmander's condition.



Beyond Hörmander's Condition

For *ensemble controllability*, we need a stronger property than Hörmander's condition.

Lie Algebra Strong Approximating Property (LASAP)

[Agrachev & Sarychev, '22] For every $Y \in \text{Vect}(M)$ and every compact $K \subset M$, there exists $\delta > 0$ such that

$$\inf_{\substack{X \in \text{Lie}\{g_1, \dots, g_m\} \\ \|X\|_{C^1(K)} \leq \delta}} \|Y - X\|_{C^0(K)} = 0.$$

In this case, the control-linear system $\dot{x} = \sum_k u_k g_k(x)$ is approximately ensemble controllable.

The Agrachev–Sarychev property also yields *finite exact ensemble controllability*:

$$x^i(0) = x^i \quad \longmapsto \quad x^i(T) = y^i, \quad i = 1, \dots, N.$$

Example on $\mathbb{R}^d$:

$$f_k(x) = \partial_{x_k}, \quad g_k(x) = e^{-\frac{|x|^2}{2}} \partial_{x_k}, \quad k = 1, \dots, d.$$

Why LASAP Gives Ensemble Controllability

Task: Approximate a target diffeomorphism

$$\Psi : K_0 \longrightarrow \Psi(K_0).$$

Let $\psi_{0,t}$ be an isotopy from Id to Ψ , generated by a time-dependent vector field Y_t , and set

$$K_t := \psi_{0,t}(K_0), \quad t \in [0, T].$$

Idea: For each t , approximate Y_t on a tubular neighborhood of K_t by elements of

$$\text{Lie}\{g_1, \dots, g_m\}.$$

Lie brackets can in turn be realized by short *concatenations of flows* along $g_1, \dots, g_m$.

$$Y_t \text{ on } K_t \approx X_t \in \text{Lie}\{g_1, \dots, g_m\} \approx \text{controlled concatenations}$$

Thus, on a short interval $(t - \varepsilon, t + \varepsilon)$, we choose controls $\tilde{u}_k(s)$ so that $\sum_{k=1}^m \tilde{u}_k(s) g_k$ *emulates the flow of Y_t near K_t* . Concatenating these local approximations yields

$$\Phi_T^u \approx \Psi \quad \text{on } K_0.$$

Main Result: Ensemble Controllability

Consider the *multi-head* control-linear Transformer $\left(V(t) = \sum_{k,l,m=1} u_{klm}(t) e_k \right)$

$$\dot{x}_i(t) = \sum_{k,l,m=1}^d u_{klm}(t) (M_i^\beta(X))_k R_{lm}(x_i(t)), \quad \text{with } R_{lm}(x) = x_l \frac{\partial}{\partial x_m} - x_m \frac{\partial}{\partial x_l}. \quad (3)$$

where $(M_i^\beta(X))_k := \langle M_i^\beta(X), e_k \rangle$ with $(e_k)_{k=1}^d$ the canonical basis of $\mathbb{R}^d$.

Theorem (Caillau-H.-Rifford '26)

Let $d \geq 3$ and $N \geq 2$. For every $X, Y \in M^N$, there exists $\beta_0 > 0$ such that, for every $\beta \geq \beta_0$ and every $T > 0$, one can find piecewise-constant controls u_{klm} steering

$$X \longrightarrow Y \quad \text{at time } T.$$

Remark

For $\beta = 0$, $M_i^0(X) = M_j^0(X)$, $\forall i \neq j$, so the system (3) is *not ensemble controllable*.

Proof Sketch I: Zero-Temperature Limit

As $\beta \rightarrow +\infty$, $M_i^\beta(X) \rightarrow x_i$, the system (3) converges to

$$\dot{x}_i = \sum_{k,l,m=1}^d u_{klm}(t) (x_i)_k R_{lm}(x_i).$$

Set $\mathfrak{L} = \text{Lie}\{x_k R_{lm} : k, l, m = 1, \dots, d\}$. The key Lie-bracket relations generate the conformal algebra $\mathfrak{so}(d, 1)$:

$$\nabla_{\mathbb{S}^{d-1}} x_k = \sum_{m=1}^d x_m R_{mk}, \quad [\nabla x_l, \nabla x_m] = R_{lm}.$$

By Lukatskii's maximality theorem,

$$\mathfrak{so}(d, 1) \subsetneq \mathfrak{L} \implies \mathfrak{L} = \{\text{polynomial vector fields on } \mathbb{S}^{d-1}\}.$$

Hence the generators satisfy LASAP, and the limit system is finite exact ensemble controllable.

Proof Sketch II: Large Finite β

On every compact $K \subset M^N$,

$$G_{klm}^\beta \longrightarrow G_{klm}^\infty \quad \text{in } C^r(K),$$

where

$$G_{klm}^\beta(X) = ((M_i^\beta(X))_k R_{lm}(x_i))_{i=1}^N.$$

Lemma

For every compact $K \subset M^N$, there exists $\beta_K > 0$ such that for all $\beta \geq \beta_K$ and $X \in K$,

$$\text{Lie}\{G_{klm}^\beta\}(X) = \text{Lie}\{G_{klm}^\infty\}(X) = T_X M^N.$$

Given $X, Y \in M^N$, choose a continuous path

$$\gamma : [0, 1] \rightarrow M^N, \quad \gamma(0) = X, \quad \gamma(1) = Y.$$

Applying the lemma to the compact set $K = \gamma([0, 1])$ yields controllability for all sufficiently large β .

Arbitrary Temperature Regime

For arbitrary $\beta > 0$, ensemble controllability remains open.

The key observation is the *separation property* of the attention

$$x_i \neq x_j \implies M_i^\beta(X) \neq M_j^\beta(X),$$

which allows interpolation polynomials as particle selectors

$$p_i(M_j^\beta(X)) = \delta_{ij}.$$

If such selectors can be generated by Lie brackets, then one obtains vector fields

$$(0, \dots, 0, R_{lm}(x_i), 0, \dots, 0),$$

and therefore independent control of each particle.

Alternative route: augment the system with $z_i \in \mathbb{R}$,

$$\begin{cases} \dot{x}_i = z_i \sum_{k,m} w_{km}(t) R_{km}(x_i), \\ \dot{z}_i = c_0(t) + c_2(t) z_i^2 + \sum_{l=1}^d u_l(t) (M_i^\beta(X))_l. \end{cases}$$

Numerical Experiments: Setup

We discretize the controlled dynamics with a projected Euler scheme:

$$\tilde{x}_i^{(l)} = x_i^{(l-1)} + \text{dt} \sum_{k,p,q=1}^d u_{kpq}^{(l)} (M_i^\beta(X^{(l-1)}))_k R_{pq}(x_i^{(l-1)}),$$

$$x_i^{(l)} = \frac{\tilde{x}_i^{(l)}}{\|\tilde{x}_i^{(l)}\|}, \quad l = 1, \dots, L.$$

Initial and target configurations $X^0 = (x_i^0)_{i=1}^N$ and $Y = (y_i)_{i=1}^N$ are sampled randomly on $\mathbb{S}^{d-1}$.

The controls are trained with Adam by minimizing

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N (1 - \langle x_i^{(L)}, y_i \rangle) + \lambda \sum_{l=1}^L \|u^{(l)}\|^2$$

Terminal angular error + L^2 control regularization.

Numerical Experiments: Results

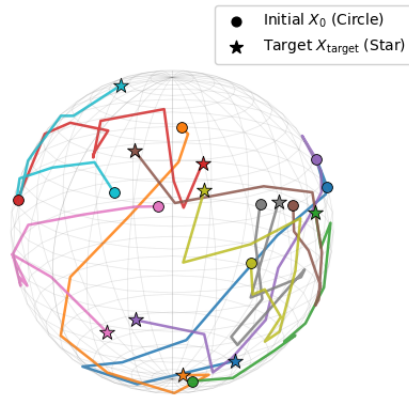
Average terminal angular error

N	β	L	$\mathbb{S}^2$ Error ($^\circ$)	$\mathbb{S}^3$ Error ($^\circ$)
10	1	3	$2.71 \cdot 10^0$	$6.24 \cdot 10^{-4}$
10	100	3	$2.70 \cdot 10^{-2}$	$5.69 \cdot 10^{-5}$
10	1	10	$1.97 \cdot 10^{-4}$	$8.21 \cdot 10^{-5}$
10	100	10	$3.81 \cdot 10^{-7}$	$1.99 \cdot 10^{-5}$
30	1	3	$2.32 \cdot 10^1$	$5.99 \cdot 10^0$
30	100	3	$1.63 \cdot 10^1$	$2.83 \cdot 10^0$
30	1	10	$2.66 \cdot 10^0$	$8.21 \cdot 10^{-3}$
30	100	10	$3.32 \cdot 10^{-1}$	$1.38 \cdot 10^{-5}$

Observations:

- Larger β generally yields lower error.
- Larger ensembles benefit from greater depth.

Sub-Riemannian Particle Steering on $\mathbb{S}^2$ ($N = 10$)



- **Measure-space controllability / Continuous Normalizing Flows.** Extend the finite-particle results to $\mathcal{P}(\mathbb{S}^{d-1})$ through

$$\partial_t \mu_t + \operatorname{div}_{\mathbb{S}^{d-1}}(\mu_t V^\beta[\mu_t]) = 0,$$

and steer a source measure μ to a target measure ν .

- **Reducing the number of controls.** The current construction uses $O(d^3)$ controls. Can one find a smaller family of vector fields still satisfying LASAP?
- **Long-time dynamics.** Understand oversmoothing, clustering, and metastability of the attention dynamics.

References :

- Agrachev, A., & Sarychev, A. (2022). Control on the manifolds of mappings with a view to the deep learning. *Journal of Dynamical and Control Systems*, 28(4), 989-1008.
- Geshkovski, B., Letrouit, C., Polyanskiy, Y., & Rigollet, P. (2025). A mathematical perspective on transformers. *Bulletin (New Series) of the American Mathematical Society*, 62(3), 427.
- Scagliotti, A., & Farinelli, S. (2025). Normalizing flows as approximations of optimal transport maps via linear-control neural ODEs. *Nonlinear Analysis*, 257, 113811.

Thank you for your **Attention** !