Training-Induced Escape from Token Clustering in a Mean-Field Formulation of Transformers

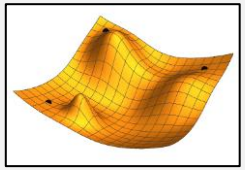
Noboru Isobe (RIKEN), Daisuke Inoue (Imperial College London),
Masaaki Imaizumi (UTokyo/RIKEN/KyotoU)

2-page summary：Math for inference in Trained TFs

Intro.: dynamical systems arising in deep learning

e.g. learning (train, optimization) dynamics

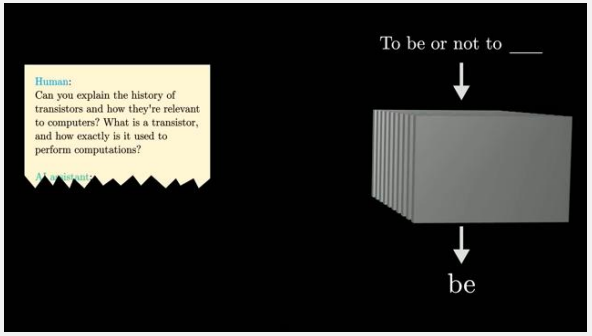
$$\theta_{m+1} = \theta_m - \eta \nabla J(\theta_m), \quad J(\theta) = \sum_{(u,y)} \mathrm{Loss}(\mathrm{DNN}(u,\theta), y)$$



Today's topic: *Inference* dynamics in neural nets

Specifically, our goal is to develop a mathematical understanding of the dynamics *in* **Transformers** (TFs).

Rough view



Cited from 3blue1brown

DNN（ResNet [He et al. 2016].）

$$u \mapsto x_1 \mapsto x_2 \mapsto x_3 \mapsto x_4 \mapsto x_5 \mapsto y =: \mathrm{DNN}(u,\theta)$$

$$x_{n+1} = x_n + v(x_n, \theta_n)$$

u; input; seq. of vectors, $(x_n)_{n=1}^L$; latent vectors

θ; trainable parameters, v; Neural net; nonlinear parametrized function

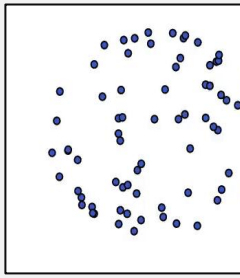
2-page summary：Math for inference in Trained TFs

Previous research [Sander+22][Geshkovski+25]... :

$$u \quad x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5$$
$$x_{n+1} = x_n + v(x_n, \theta_n)$$

TFs are modeled by evolution of *probability distributions*.

✓ Under a simple choice of trainable param. θ,
the latent seq. tends to "cluster"(or a Dirac delta δ)



➢Provides one way of understanding implicit bias in TFs.

◆ In practice, θ is optimized & the latent $\not\to \delta$ (detailed later)

Our setting & result [I.-Inoue-Imaizumi,26] :

See the impact of training on inference dynamics through *a turnpike analysis* of the optimal control prob.

$$\underset{W \in L^2(0,T)}{\text{minimize}}\, J(W) := \ell\left(\mu_T^W\right) + \frac{\lambda_{\text{reg}}}{2}\|W\|_{L^2(0,T)}^2,$$

$$\text{subject to} \begin{cases} \partial_t \mu + \operatorname{div}\left(\mu P^{\perp}\left(-\nabla \frac{\delta E_{\varepsilon,A}}{\delta \mu} + u_{W_t}\right)\right) = 0, \\ E_{\varepsilon,A}(\mu) := \varepsilon\, \mathrm{Ent}(\mu) - \frac{1}{2}\iint \mathrm{e}^{\langle x, Ay\rangle} \mathrm{d}\mu(x)\mathrm{d}\mu(y). \end{cases}$$

✓ ***Escape* from clustering** near the terminal layers $\approx T$

Contents

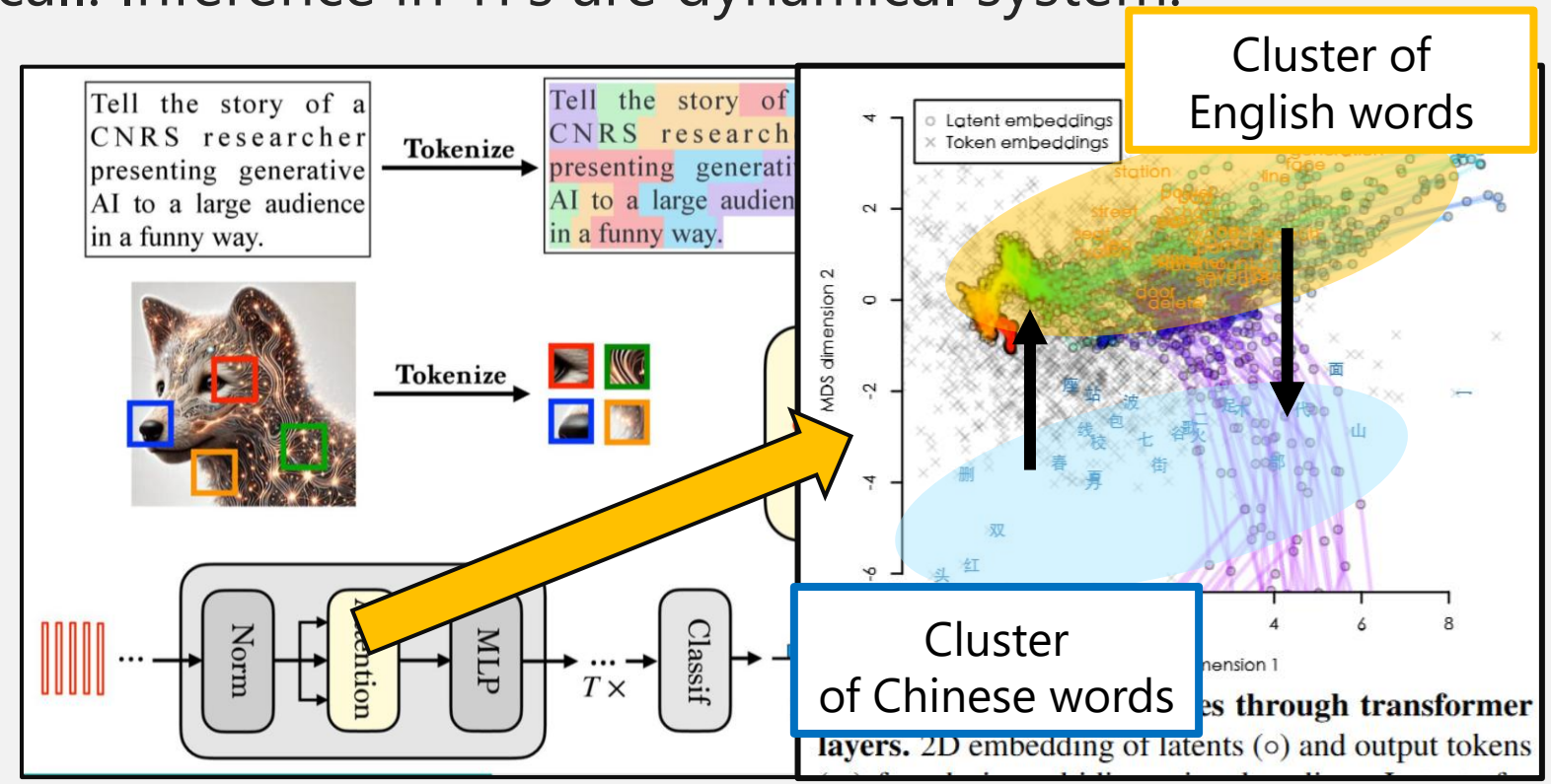
I Backgrounds
- Empirical results about inference in TFs
- TF as particle systems, and mean-field models

II Problem setting
- Noisy Transformer with trainable linear MLPs
- Training as mean-filed optimal control problems

III Mathematical result
- Exponential turnpike theorem for optimal TFs

Background: Clustering in Transformers

Recall: Inference in TFs are dynamical system:



Cited from the Peyré's slide

- Empirical results in [Wendler+,ACL24] shows that LLM (Llama2) "thinks in English" despite Chinese inputs.

➢ Aim to clarify the mechanism mathematically

Background: Transformer architecture

Review (Transformer: TF)

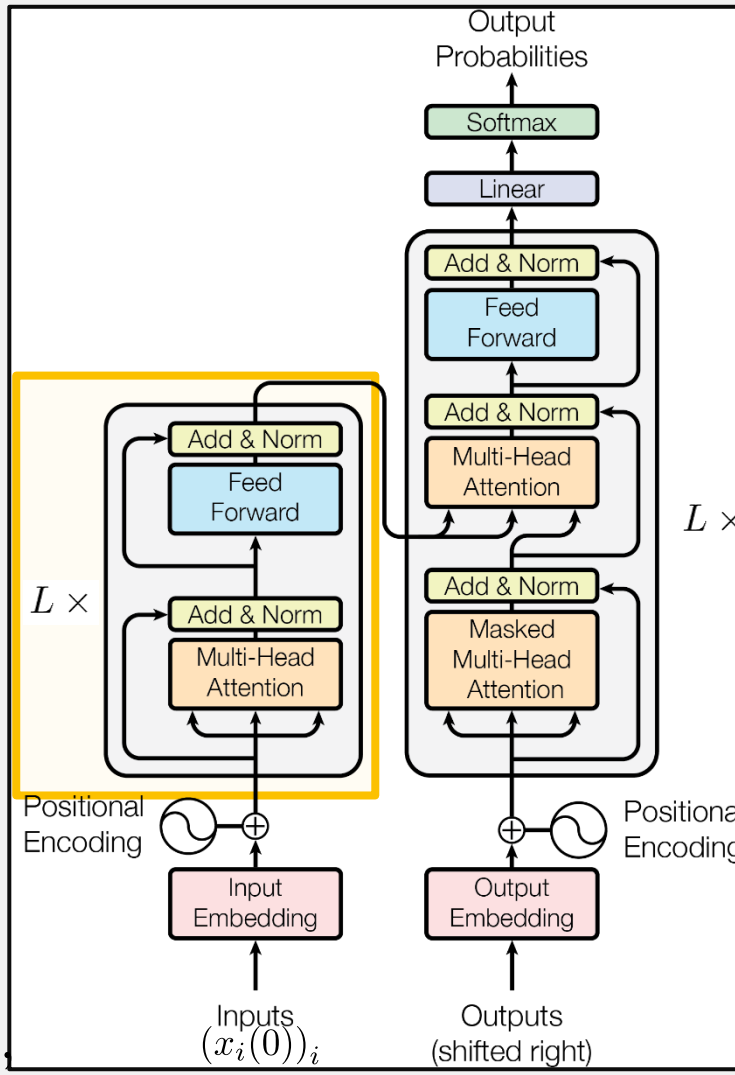
TF is a parameterized function that returns *a sequence* from a *sequence*, that is…

- Input: $(x_i(0))_{i=1}^N \in (\mathbb{R}^d)^N$

  N can be changed.

- For each layer $l \in \{1, \ldots, L\}$ recursively define $(x_i(l))_{i=1}^N$ as follows (cont'd):

$$x_i(l+1) = \mathcal{N}\left(x_i(l) + \mathrm{FFN}(x_i(l)) + \sum_{h=1}^{H} \mathrm{Attn}_h((x_i(l))_i) \right), \quad l = 0, \ldots, L-1,$$



[Vaswani+,NIPS`17]

Review (Transformer: TF)

$$(x_i(0))_{i=1}^N \in (\mathbb{R}^d)^N$$

$$x_i(l+1) = \mathcal{N}\left(x_i(l) + \mathrm{FFN}(x_i(l)) + \sum_{h=1}^{H} \mathrm{Attn}_h((x_i(l))_i) \right), \quad l = 0, \dots, L-1,$$

where

- $\mathcal{N}\colon \mathbb{R}^d \longrightarrow \mathbb{S}^{d-1}$; Projection onto the sphere in this talk
- $\mathrm{FFN}\colon \mathbb{R}^d \longrightarrow \mathbb{R}^d$; Feedforward net e.g. $F(x) = W\,\mathrm{Activation}(Ax+b)$ $(W, A, b;$ learnable param.$)$
- $\mathrm{Attn}_h\colon (\mathbb{R}^d)^N \longrightarrow \mathbb{R}^d$; *Self-Attention* given by

$$\mathrm{Attn}_h((x_i(l))_i) := \frac{1}{Z} \sum_{j=1}^{N} \exp\left(\beta \left\langle Q_l^h x_i(l), K_l^h x_j(l) \right\rangle\right) V_l^h x_j(l)$$

$$\beta \approx d^{-1/2} \quad Z := \sum\nolimits_j \mathrm{e}^{\beta \left\langle Q_l^h x_i(l), K_l^h x_j(l) \right\rangle}$$

$(Q, K, V;$ learnable matrices$)$

Theoretical framework: Continuous limit of TFs

Let us focus on the case only with Single Attention:

$$x_i(l+1) = \mathcal{N}\left(x_i(l) + \mathrm{FFN}(x_i(l)) + \sum_{h=1}^{H} \mathrm{Attn}_h((x_i(l))_i) \right)$$

The discrete-time dynamics remain somewhat complex.

Another approach inspired by [E, `17]:

- Take **a "continuous limit" w.r.t. the layers** of DNNs and study dynamical properties of the differential equation. For example…

DNN (ResNet [He et al. 2016].)

$$u \mapsto x_1 \mapsto x_2 \mapsto x_3 \mapsto x_4 \mapsto x_5 \mapsto y$$

$$x_{n+1} = x_n + v(x_n, \theta_n)$$

ODE-Net (Neural ODE [Chen et al. '18])

$$u \mapsto x(t), 0 \le t \le 1 \mapsto y$$

$$\dot{x}(t) = v\left(x(t), \theta(t)\right)$$

Theoretical framework: Continuous limit of TFs

Let us move to our "attention" from

$$x_i(l+1) = \mathcal{N}\left(x_i(l) + \mathrm{Attn}((x_i(l))_i) \right)$$

to the following continuous dynamics [Geshkowski+,AMS`25]:

$$\dot{x}_i(t) = P^{\perp}_{x_i(t)} \left(\sum_{j=1}^{N} \frac{\mathrm{e}^{\beta \langle Q x_i(t), K x_j(t) \rangle}}{Z_{\beta,i}(t)} V x_j(t) \right), \quad i \in \{1, \dots, N\}, \quad t > 0,$$

where

- $x_i \colon [0, +\infty) \longrightarrow \mathbb{S}^{d-1}$ $(i \in [N])$; conti. traj. of tokens
  - ✓ Time t corresponds to the layer l
  - ✓ Consider ideal "infinite numbers of layer" $(t \to \infty)$
- $P^{\perp}_x := I - xx^{\top} \in \mathbb{R}^{d \times d}$ for $x \in \mathbb{S}^{d-1}$; proj. onto $T_x \mathbb{S}^{d-1}$
- $Q, K, V \in \mathbb{R}^{d \times d}$; QKV weights ($t$-independent for simplicity)
- $\beta > 0$, $Z_{\beta,i}(t) := \sum_j \mathrm{e}^{\beta \langle Q x_i(t), K x_j(t) \rangle}$; partition function

Theoretical framework: Mean-field view of TFs

$$\dot{x}_i(t) = P^{\perp}_{x_i(t)} \left(\sum_{j=1}^{N} \frac{\mathrm{e}^{\beta \langle Q x_i(t), K x_j(t) \rangle}}{Z_{\beta,i}(t)} V x_j(t) \right), \quad i \in \{1, \dots, N\}, \quad t > 0,$$

Remark: The above is permutation invariant w.r.t. i .

So let us denote $\mu(t)$ by the empirical (probability) distribution consisting of a Dirac delta as $\mu(t) := \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i(t)} \in \mathcal{P}(\mathbb{S}^{d-1})$ (Borel prob. measures)

Then we instead have another formulation of TF as

$$\begin{cases} \partial_t \mu + \operatorname{div} \left(\chi_\beta[\mu] \mu \right) = 0 & \text{on } \mathbb{R}_{>0} \times \mathbb{S}^{d-1} \\ \mu|_{t=0} = \mu(0) := \frac{1}{N} \sum_i \delta_{x_i(0)} & \text{on } \mathbb{S}^{d-1} \end{cases},$$

$$\chi_\beta[\mu](x) := P^{\perp}_x \int_{\mathbb{S}^{d-1}} \frac{\mathrm{e}^{\beta \langle Qx, Ky \rangle}}{Z_{\beta,\mu}(x)} V y \, \mathrm{d}\mu(y), \quad Z_{\beta,\mu}(x) := \int_{\mathbb{S}^{d-1}} \mathrm{e}^{\beta \langle Qx, Ky \rangle} \mathrm{d}\mu(y)$$

- Nonlinear & non-local PDE on the high-dim sphere

Previous results: emergence of “clusters”

$$\begin{cases} \partial_t \mu + \operatorname{div}(\chi_\beta[\mu]\mu) = 0 & \text{on } \mathbb{R}_{>0} \times \mathbb{S}^{d-1} \\ \mu|_{t=0} = \mu(0) := \frac{1}{N}\sum_i \delta_{x_i(0)} & \text{on } \mathbb{S}^{d-1} \end{cases},$$

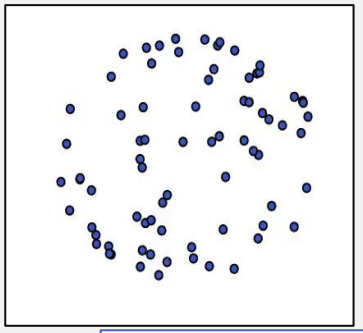
$$\chi_\beta[\mu](x) := P_x^\perp \int_{\mathbb{S}^{d-1}} \frac{\mathrm{e}^{\beta\langle Qx, Ky\rangle}}{Z_{\beta,\mu}(x)} Vy \, \mathrm{d}\mu(y), \quad Z_{\beta,\mu}(x) := \int_{\mathbb{S}^{d-1}} \mathrm{e}^{\beta\langle Qx, Ky\rangle} \mathrm{d}\mu(y)$$

Previous works [Geshkowski+23,24,25, Bruno+,25,...]:

Fix QKVs as $Q = K = V = I$ (and omit Z).

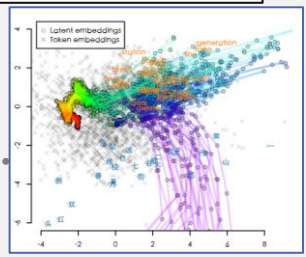
- leads the (Wasserstein) gradient-flow structure formally given by $\dot{\mu} = -\operatorname{grad}\mathcal{E}(\mu)$.

➢ The set of minimizers of $\mathcal{E}$ consists of Dirac deltas, so the flow is expected to be “clusters”



These results are interesting mathematically.

But there’re difference between Theory and Practice.



Contents

I Backgrounds
- Empirical results about inference in TFs
- TF as particle systems

II Problem setting
- Noisy Transformer with trainable linear MLPs
- Training as mean-filed optimal control problems

III Mathematical result
- Exponential turnpike theorem for optimal TFs

Our model: Noisy Transformer w/ trainable FFNs

By technical reasons, we instead consider another dynamical model of Transformer w/ *trainable* params W in Feedforward nets:

$$\begin{cases} \partial_t \mu + \operatorname{div}\left(P_x^{\perp}(\chi_A[\mu] + \mathrm{FFN})\mu\right) = \varepsilon\Delta\mu & \text{on } \mathbb{R}_{\geq 0} \times \mathbb{S}^{d-1} \\ \mu|_{t=0} = \mu(0) & \text{on } \mathbb{S}^{d-1} \end{cases}, \quad \chi_A[\mu](x) := \int \mathrm{e}^{\langle x, Ay\rangle} Ay \, \mathrm{d}\mu(y)$$

$$\mathrm{FFN}(x) = u_{W_t}(x) = W_t\sigma(x), W_t \in \mathbb{R}^{d\times p}, \sigma\colon \mathbb{R}^d \to \mathbb{R}^p; \mathrm{Lip}.$$

That is, we add or modify

- Noise (or Laplacian) term $\varepsilon\Delta\mu$ $(\varepsilon > 0)$ for regularity
- Symmetric Attention w/ $A(= Q^\top K)) \in \mathbb{R}^{d\times d}, V = Q^\top K$
- Linear FFN is sufficient to prove the existence of the minimizer introduced in the next page;

Our prob.: Math. model of trained TFs as opt. control

Consider the following mean-field optimal control problem:

$$\underset{W \in L^2(0,T;\mathbb{R}^{d\times p})}{\text{minimize}} \left\{ \ell(\mu_T^W) + \frac{\lambda_{\text{reg}}}{2} \|W\|_{L^2}^2 \right\},$$

where ℓ is a non-negative "smooth" function on prob. meas's

e.g. (softmax)
$-\sum_{w \in \text{Vocab.}} q_{\text{techer}}(w) \log \left(\int \text{softmax}(\text{LN}(x))_w \, \mathrm{d}\mu(x) \right).$

$(\mu_t^W) \in C([0,T]; \mathcal{P}(\mathbb{S}^{d-1}))$ e following equation

$$\begin{cases} \partial_t \mu_t^W + \text{div}\left(\mu_t^W P_x^\perp \left(W_t \sigma + \chi_A[\mu_t^W]\right)\right) = \varepsilon \Delta \mu_t^W & \text{in } \mathbb{S}^{d-1} \times (0,T), \\ \mu_t^W\big|_{t=0} = \mu_0 & \text{on } \mathbb{S}^{d-1}. \end{cases}$$

- $T, \lambda_{\text{reg}}, \varepsilon > 0$
- $P_x^\perp = I - xx^\top \in \mathbb{R}^{d\times d}$ $(x \in \mathbb{S}^{d-1})$: proj. onto $T_x\mathbb{S}^{d-1}$
- $\sigma : \mathbb{S}^{d-1} \to \mathbb{R}^d$: Lipschitz
- $A \in \mathbb{R}^{d\times d}$: symmetric positive definite (SPD), and

$$\chi_A[\mu](x) = \int_{\mathbb{S}^{d-1}} \mathrm{e}^{\langle x, Ay \rangle} Ay \, \mathrm{d}\mu(y)$$

Contents

I Backgrounds
- Empirical results about inference in TFs
- TF as particle systems

II Problem setting
- Noisy Transformer with trainable linear MLPs
- Training as mean-filed optimal control problems

III Mathematical result
- Exponential turnpike theorem for optimal TFs

Observation: Gradient flow structure in TF

Consider an optimal solution of the following:

$$\underset{W\in L^2(0,T;\mathbb{R}^{d\times p})}{\text{minimize}} \quad \ell(\mu_T^W) + \frac{\lambda}{2}\int_0^T \|W_t\|_{\text{frob}}^2 \,\mathrm{d}t$$

$$\text{subject to} \quad \begin{cases} \partial_t \mu^W + \operatorname{div}\big(P_x^{\perp}(\chi_A[\mu^W] + W\sigma)\mu^W\big) = \varepsilon\Delta\mu^W & \text{on } \mathbb{R}_{\geq 0}\times\mathbb{S}^{d-1}, \\ \mu^W|_{t=0} = \mu(0) & \text{on } \mathbb{S}^{d-1}. \end{cases}$$

Observation (gradient flow structure): $\chi_A[\mu](x) = \int_{\mathbb{S}^{d-1}} \mathrm{e}^{\langle x, Ay\rangle} Ay \,\mathrm{d}\mu(y)$

We can rewrite the eq. as

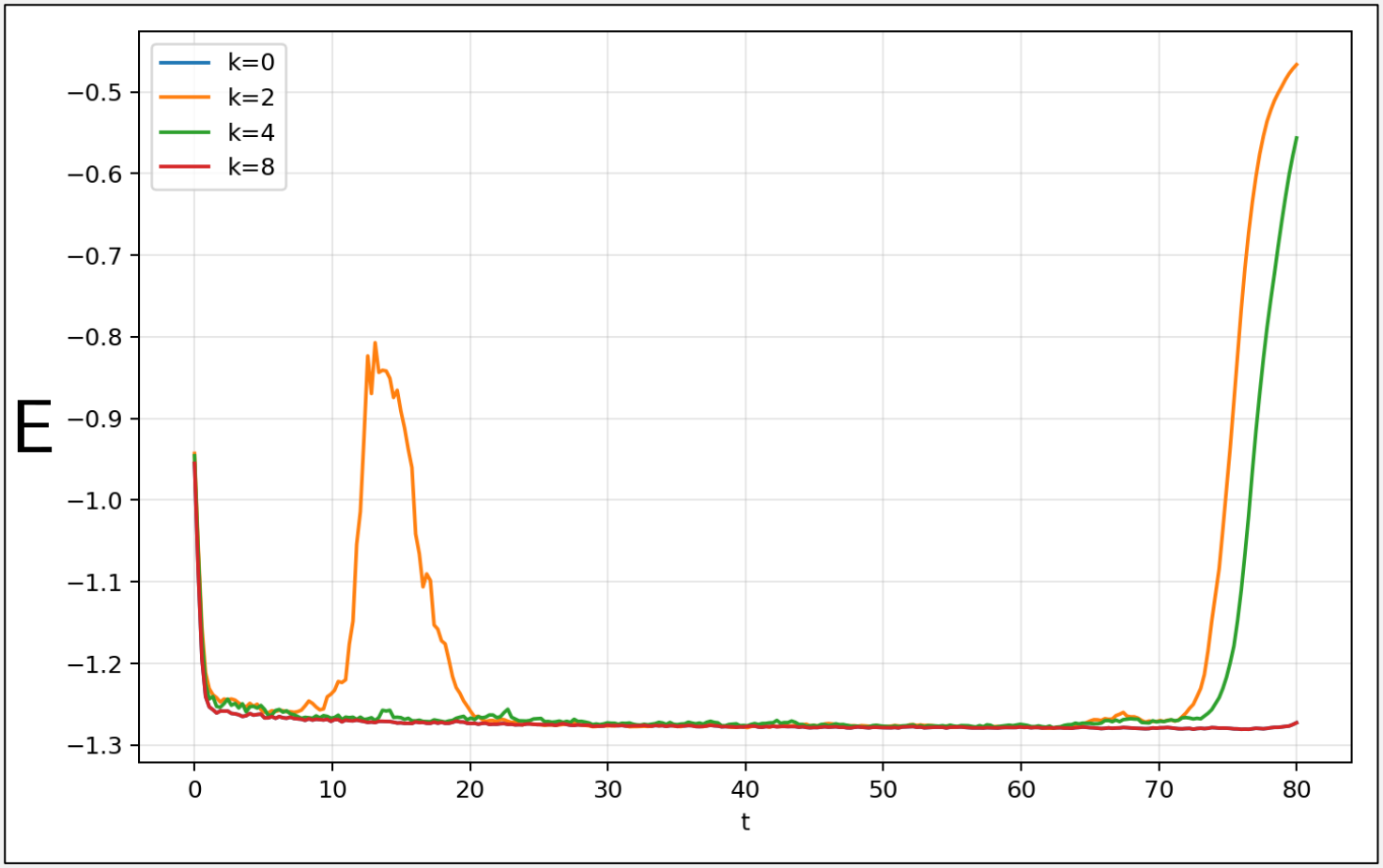
$$\partial_t \mu^W + \operatorname{div}\left(\mu^W v_t^W\right) = 0, v_t^W(x) = P_x\left(-\nabla\frac{\delta E_{\varepsilon,A}}{\delta\mu} + W\sigma\right),$$

$$E_{\varepsilon,A}(\mu) := \varepsilon\mathrm{Ent}(\mu) - \frac{1}{2}\iint \mathrm{e}^{\langle x,Ay\rangle}\mathrm{d}(\mu\otimes\mu)(x,y).$$

- If $W \approx 0$, this flow "minimizes" the energy E
  - $|W\sigma(x)| \lesssim \|W\|_{\text{frob}}$ (activation σ is b'dd)

Informal discussion: Numerical observation E

Numerical experiments show that the optimal trajectory of $\mathrm{E}(\mu_t^{W^*})$ seems to have "U-shaped";



Result：Optimized traj.s almost remain stationary

Set $\Delta_{\varepsilon,A}(\mu) := E_{\varepsilon,A}(\mu) - \min E_{\varepsilon,A}$, $\Delta_0 := \frac{(\sinh \lambda_1) \wedge (\mathrm{e}^{\lambda_1} - \mathrm{e}^{\lambda_2})}{2}$

Assumption：

1. The SPD A has a simple max. eigen value λ_1.
2. $0 < \varepsilon \ll 1$.
3. Input μ_0 and $\operatorname{osc} \ell = \max \ell - \min \ell$ satisfy
$$(\Delta_{\varepsilon,A}(\mu_0) + \tfrac{\|\sigma\|_\infty^2}{\lambda_{\mathrm{reg}}} \operatorname{osc} \ell) \vee (\tfrac{4\|\sigma\|_\infty^2}{\lambda_{\mathrm{reg}}} \operatorname{osc} \ell) < \Delta_0$$

Theorem (exp. turnpike bound)： Under Assumptions, $\exists\, T^*, C, a > 0$, $\forall\, T > T^*$, $\forall\, (\mu^*, W^*)$: minimizer, we have

$$E_{\varepsilon,A}(\mu_t^*) - \min E_{\varepsilon,A} \le C\left(\mathrm{e}^{-at} + \mathrm{e}^{-a(T-t)}\right), \quad t \in [0,T].$$

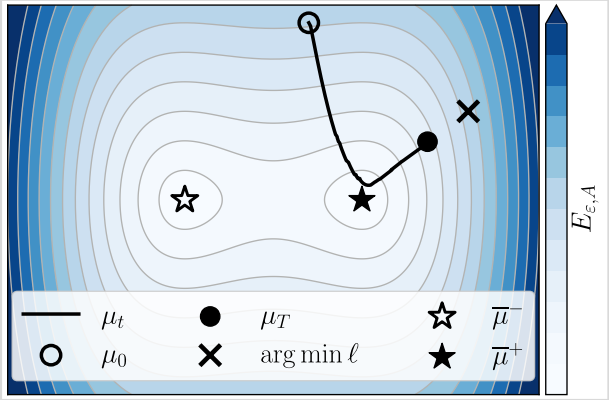
$$\underset{W \in L^2(0,T)}{\text{minimize}}\, J(W) := \ell(\mu_T^W) + \frac{\lambda_{\mathrm{reg}}}{2}\|W\|_{L^2(0,T)}^2, \text{ subject to } \begin{cases} \partial_t \mu + \operatorname{div}\left(\mu P^\perp\left(-\nabla \frac{\delta E_{\varepsilon,A}}{\delta \mu} + u_{W_t}\right)\right) = 0, \\ E_{\varepsilon,A}(\mu) := \varepsilon \operatorname{Ent}(\mu) - \frac{1}{2}\iint \mathrm{e}^{\langle x, Ay\rangle} \mathrm{d}\mu(x)\mathrm{d}\mu(y). \end{cases}$$

Key Lemma: positivity of the formal Hessian

$$E_{\varepsilon,A}\left(\mu_t^*\right) - \min E_{\varepsilon,A} \leq C\left(\mathrm{e}^{-at} + \mathrm{e}^{-a(T-t)}\right), \quad t \in [0,T].$$

Informal explanation The optimized $(\mu_t^*)_t$ moves to...

1. Decrease the energy $E_{\varepsilon,A}$,
2. Stagnate in nhd of $\operatorname{argmin} E_{\varepsilon,A}$,
   Here we use $(\Delta_{\varepsilon,A}(\mu_0) + \frac{\|\sigma\|_\infty^2}{\lambda_{\mathrm{reg}}} \operatorname{osc} \ell) \vee (\frac{4\|\sigma\|_\infty^2}{\lambda_{\mathrm{reg}}} \operatorname{osc} \ell) < \Delta_0$
3. Escape (to minimize the loss ℓ.)



The next lemma is helpful to deduce the exp. decay/growing:

Lemma (→PŁ ineq.) The Hessian H at a minimizer $\overline{\mu}$ is positive definite for $h \in L_0^2(\overline{\mu}) := \left\{h \in L^2(\overline{\mu}) \mid \int h \, \mathrm{d}\overline{\mu} = 0\right\}$ as

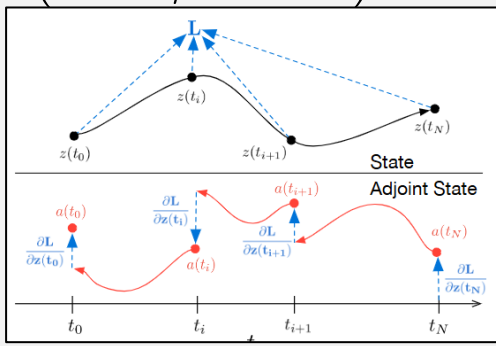
$$\langle h, Hh \rangle_{L^2(\overline{\mu})} \geq \frac{\varepsilon}{2}\left(1 - \frac{\lambda_2}{\lambda_1}\right)\|h\|_{L^2(\overline{\mu})}^2.$$

Key Lemma: positivity of the formal Hessian

$$E_{\varepsilon,A}\left(\mu_t^*\right) - \min E_{\varepsilon,A} \leq C\left(\mathrm{e}^{-at} + \mathrm{e}^{-a(T-t)}\right), \quad t \in [0,T].$$

Proof Sketch: Reflecting the structure $\partial_t \mu_t + \operatorname{div}\left(-\nabla \frac{\delta E_{\varepsilon,A}}{\delta \mu} + W\sigma\right) = 0$ the gradients ϕ via backpropagation follows →adjoint another dissipative system in *reverse* time.

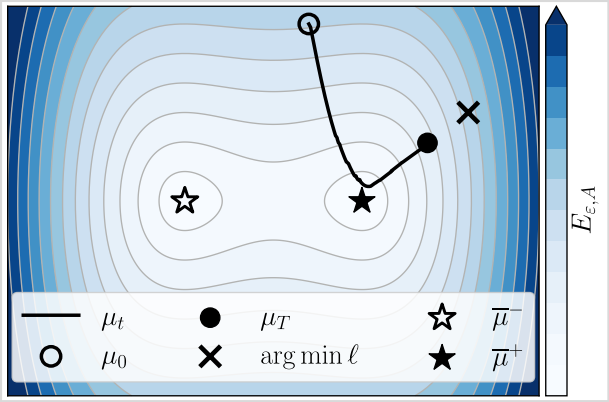
- In the space of (μ, ϕ), the traj. moves (stable/unstable) mfds around μ.



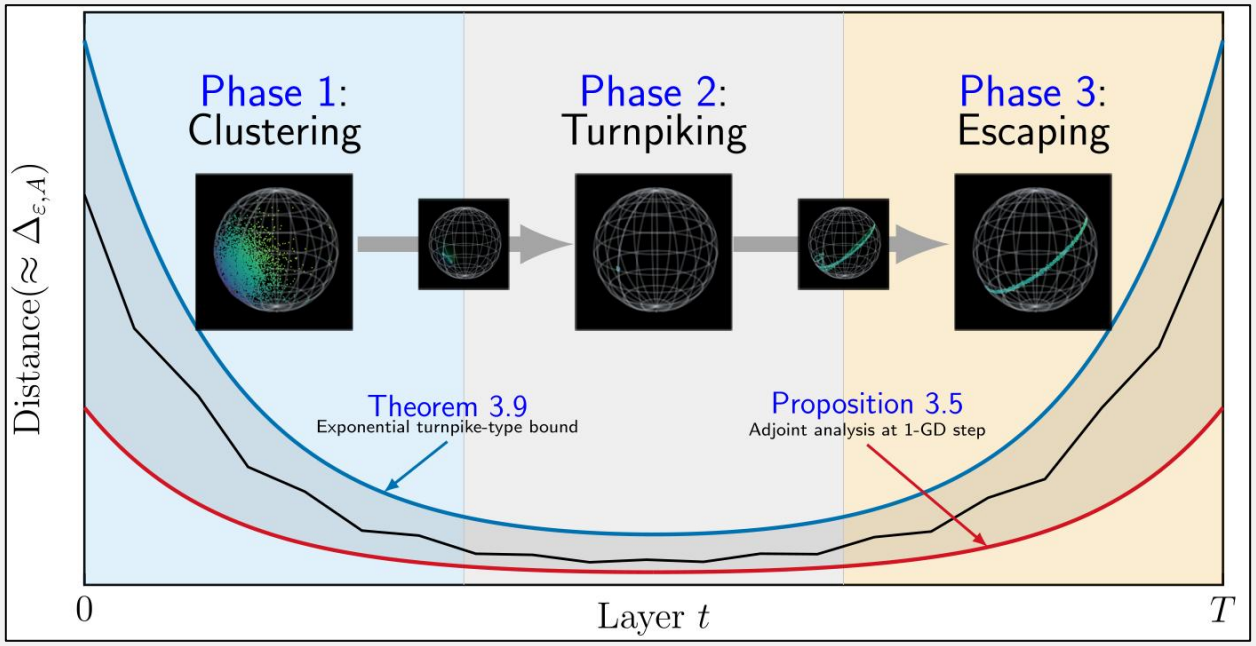
Cited from [Chen+18]

Informal explanation The optimized $(\mu_t^*)_t$ moves to...

1. Decrease the energy $E_{\varepsilon,A}$,
2. Stagnate in nhd of $\operatorname{argmin} E_{\varepsilon,A}$,
   Here we use $(\Delta_{\varepsilon,A}(\mu_0) + \frac{\|\sigma\|_\infty^2}{\lambda_{\mathrm{reg}}} \operatorname{osc} \ell) \vee (\frac{4\|\sigma\|_\infty^2}{\lambda_{\mathrm{reg}}} \operatorname{osc} \ell) < \Delta_0$
3. Escape (to minimize the loss ℓ.)



Three phases: clustering, turnpike, escape



Phase 1 · Clustering

early layers contract toward the cluster.

Phase 2 · Turnpike

a long plateau; the control is nearly idle.

Phase 3 · Escape

the loss pulls tokens off the cluster.

Remark: Jumping around final layers in TFs

Some of experimental results on mechanistic interpretability of Transformer find a jump at final layer:

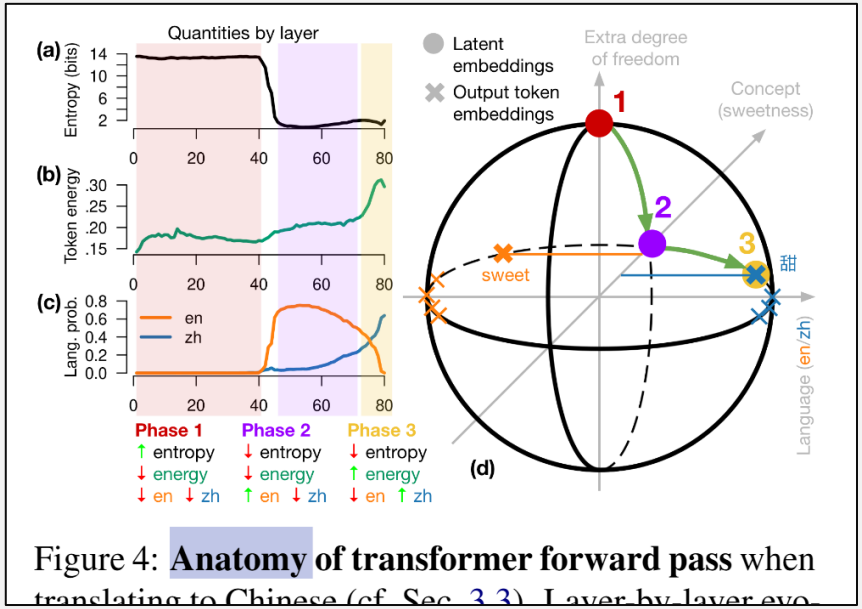
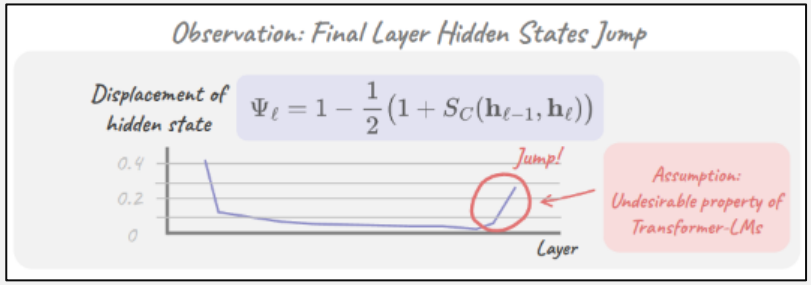


Figure 4: **Anatomy of transformer forward pass** when translating to Chinese (cf. Sec. 3.3). Layer-by-layer evo-

In Llama-2
[Wendler+24]



In Llama-3, Gemma, DeepSeek
[Shibata+26]

Conclusion & outlook

Training changes mean-field Transformer inference:

$$\text{clustering} \;\to\; \text{turnpike} \;\to\; \text{escape}$$

with an exponential turnpike bound.

Future:

- the role of more general matrix A
  - ✓ Empirically, A has complex eigenvalues and leads oscillating behavior of tokens.
- Common noise case (cf. [Koubbi+26+,Agazzi+,26+])
- ...(I believe we can consider many variants of the problem)

Thank you!

arXiv:2605.07772 · Isobe – Inoue – Imaizumi