Universal Phenomena, Irreversibility, and Thermodynamics in Deep Representation Learning

Liu Ziyin

MIT / NTT Research

2025/08/09

Liu Ziyin*, Yizhou Xu*, Isaac Chuang. *Neural Thermodynamics I: Entropic Forces in Deep and Universal Representation Learning*. arxiv 2505.12387

Irreversibility

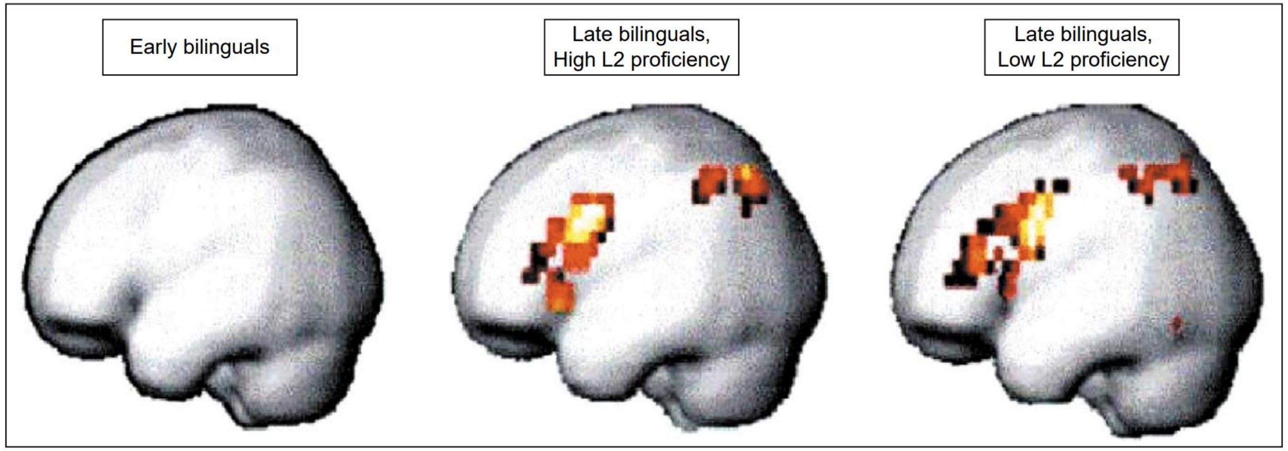
- Our world is dominated by irreversible processes
  - Time only goes one way (Arrow of time)
  - Entropy of the universe must increase (Second law of thermodynamics)

Irreversibility in the brain

- Critical period is a kind of irreversible process
- Visual deprivation experiments by Hubel and Wiesel (1963)
    - Newborn cat with one eye covered does not develop vision for that eye
    - Adult cat with one eye covered is not affected
    - A critical periods of roughly two months after birth

Irreversibility in the brain

- Critical period is a kind of irreversible process
- Language acquisition (Lenneberg 1967)
  - Natural efficiency can only be achieved when you learn language before certain age


Early bilinguals
Late bilinguals, High L2 proficiency
Late bilinguals, Low L2 proficiency

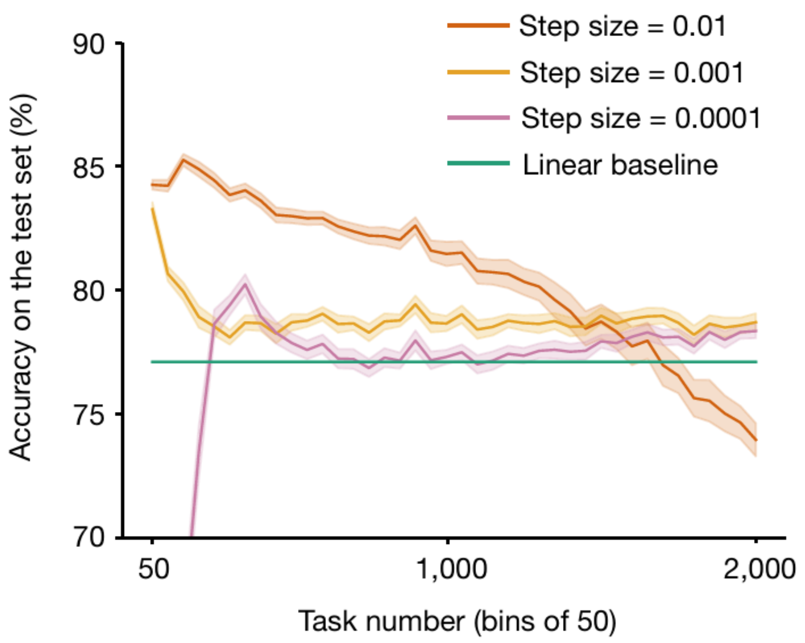
Peroni and Abutalebi (2005)

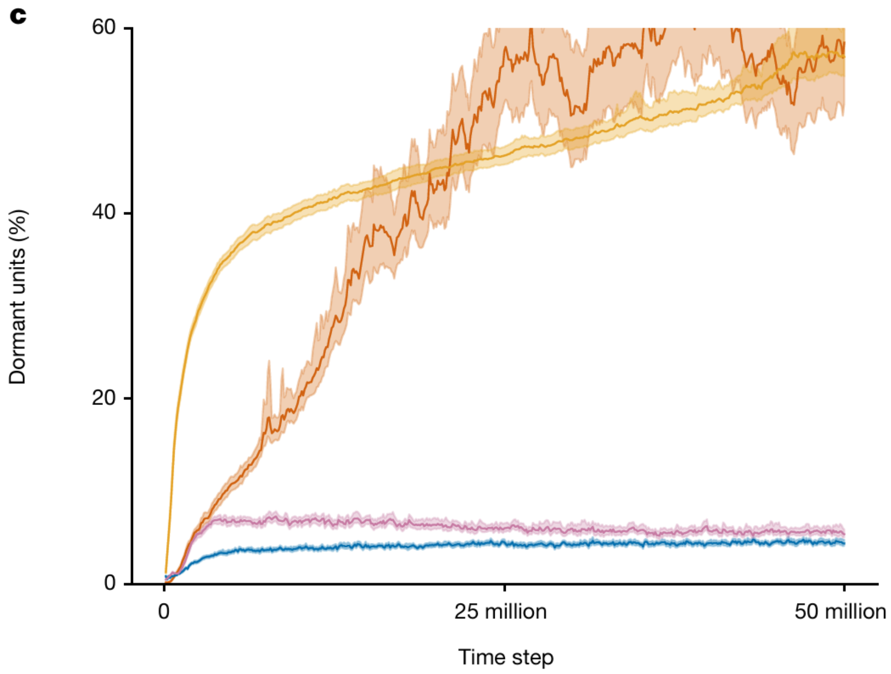
Irreversibility in AI Training

- The learning dynamics of neural networks can be divided into at least two phases (Fort et al. 2020)
  - The “**chaos transient**”: the initial few steps of training crucially determines the learning outcome; changes in the initial few steps affects the final performance the most

Irreversibility in AI Training

- Loss of plasticity (Dohare et al., 2023)
  - More and more neurons “die” during training
  - Reinforcement learning, continual learning





Irreversibility is at the core of learning.

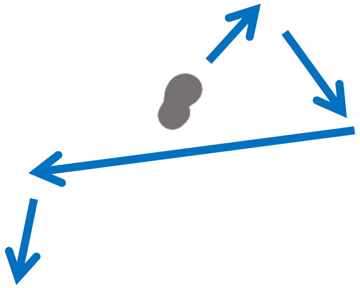
Irreversibility in Nature

- Thermal systems in nature often evolve to minimize the Free energy:

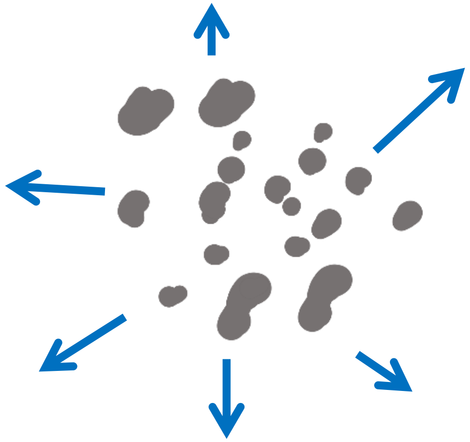
$$F = E - TS$$

- This means that the system will minimize energy while maximizing entropy
- The dynamical tendency towards maximizing the entropy can be imagined as coming from a formal "**entropic force**"
  - Leads to irreversibility
- Many phenomena in nature are due to entropic forces
  - Such as phase transitions

Entropic Force



Microscopic Random Motion



Macroscopic Flow

Free Energy in Artificial Learning

- What role does entropic force play in artificial learning?
- Entropy production must be due to irreversible processes
- Identify microscopic irreversible components of the learning dynamics and use that to define “**entropy force**”

Contents

1. Learning dynamics
2. Entropic Force Breaks Continuous Symmetries
3. Platonic Representation Hypothesis
4. Progressive Sharpening

Learning Dynamics

- In learning, we want to minimize a loss function $L(\theta)$ that is the expectation of a stochastic loss function over a large training set
$$L(\theta) = \mathbb{E}_x[\ell(x, \theta)]$$
- A magical algorithm that works very well is gradient descent (GD)
$$\Delta\theta_t = -\eta\nabla_\theta L$$
- η is the **learning rate**
  - It is unit of time
  - Its sign is the **arrow of time**
- Because L is what we will minimize, we will refer to L as the "**energy**"

Gradient Flow

- GD: $\Delta\theta_t = -\eta\nabla_\theta L$
- GF: $\frac{d}{dt}\theta = -\eta\nabla_\theta L$

Stability and Speed: **A Naïve Picture**

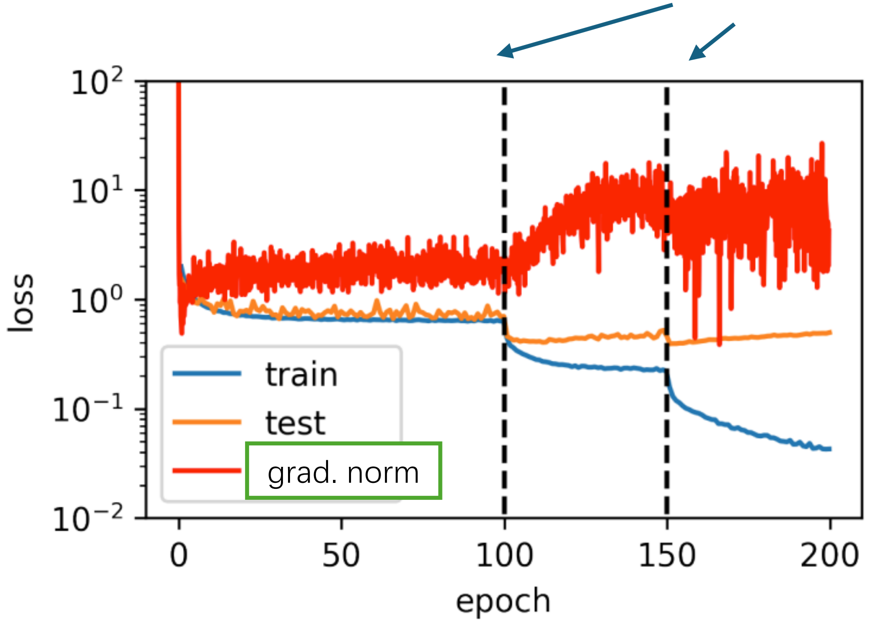
- Small η: learning is slow but stable
- Large η: fast but unstable

Stability and Speed: **A Naïve Picture**

- Small η: learning is slow but stable
- Large η: fast but unstable

Learning rate decrease

Resnet 18



Stochastic Learning Dynamics

- In learning, we want to minimize a loss function $L(\theta)$ that is the expectation of a stochastic loss function over a large training set
$$L(\theta) = \mathbb{E}_x[\ell(x, \theta)]$$
- It is impossible in reality to compute L, and so we sample an independent x at every t, and train on ℓ:
$$\Delta\theta_t = -\eta \nabla_\theta \ell(x, \theta)$$
- This is the **SGD** algorithm

Irreversibility of SGD Dynamics

- Consider running GD in a harmonic potential ($L = \frac{\theta^2}{2}$)

$$\Delta\theta_t = -\eta\theta_t$$

- If the dynamics is reversible, then θ will return to its initial position if we reverse the arrow of time
- But this is not the case for SGD:

$$\theta_1 = \theta_0 - \eta\theta_0 = (1-\eta)\theta_0$$
$$\theta_2 = \theta_1 + \eta\theta_1 = \left(1-\eta^2\right)\theta_0$$

- Reversing the time creates an error of order η^2
  - SGD is irreversible
  - **SGD can be reversible if we move very slowly** (using an infinitesimal η)

Gradient flow dynamics is always reversible

- Find $\widetilde{F}_\eta(\theta)$ such that (assuming GD)
- If

$$\Delta\theta = -\eta\nabla L$$

$$\frac{d}{dt}\tilde{\theta}_1 = -\eta\nabla\tilde{F}_\eta$$

- Then

$$\theta - \tilde{\theta} \approx 0$$

Effective Free Energy

- Because we already know that

$$F_0 = L$$

- We can expand F_η in η:

$$F_\eta = L + \eta\phi_1 + O(\eta^2)$$

- Plug into the dynamics (GD vs GF)
- We obtain a very simple form:

$$\phi_1 = \left||\nabla L|\right|^2$$

Smith et al. (2021)
“Modified loss”

Alternative derivation Irreversibility

- Find $F_\eta(\theta)$ such that (assuming GD)
- If

$$\theta_1 = \theta_0 - \eta \nabla L$$
$$\theta_2 = \theta_1 + \eta \nabla F_\eta$$

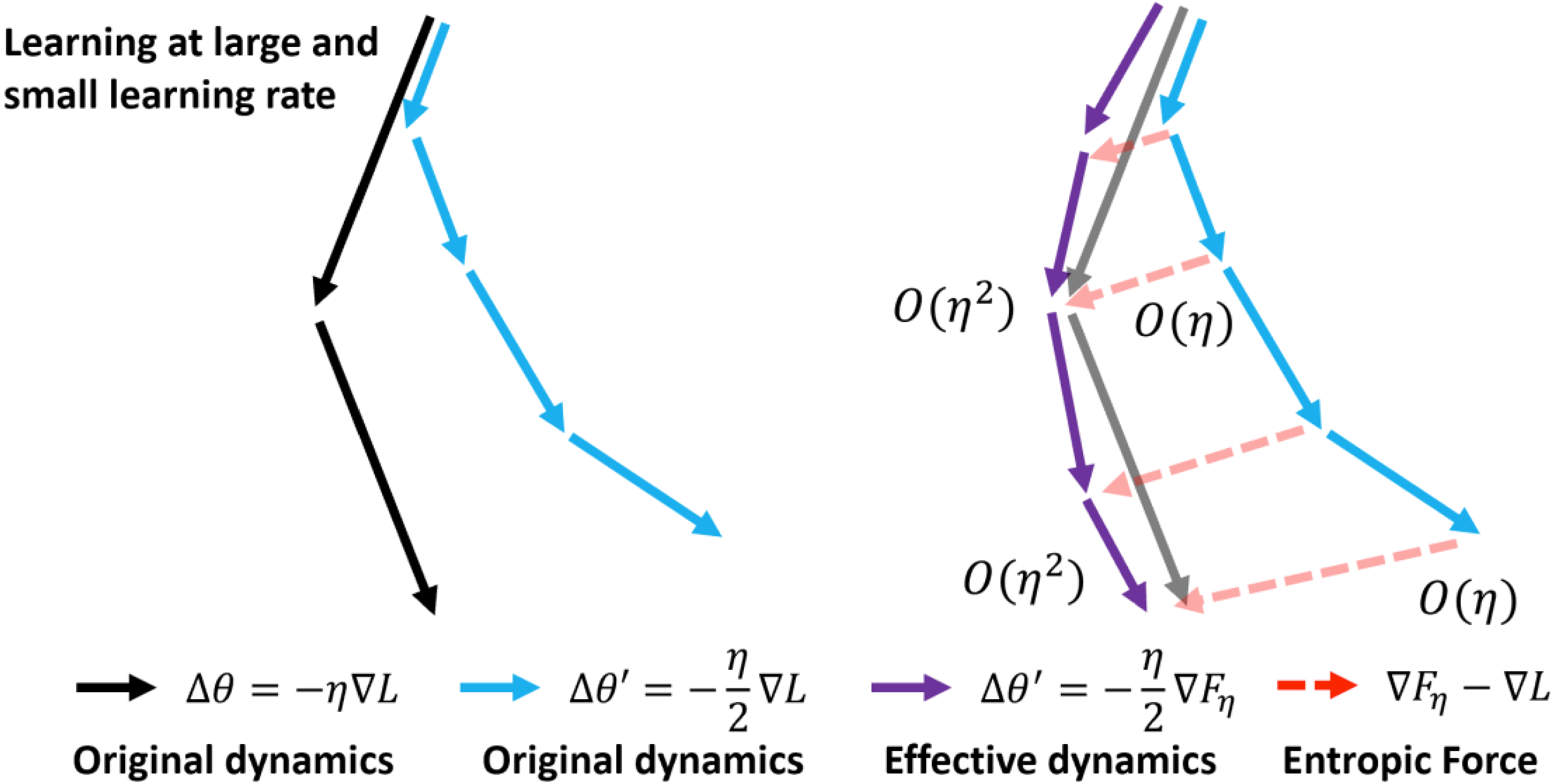
- Then

$$\theta_0 - \theta_2 \approx 0$$

$$\boxed{\widetilde{F}_\eta(\theta) = F_\eta(\theta)}$$

Irreversibility

- Making GD reversible ≈ Making GD continuous-time


Learning at large and small learning rate
$O(\eta^2)$
$O(\eta)$
$O(\eta^2)$
$O(\eta)$
$\Delta\theta = -\eta\nabla L$
Original dynamics
$\Delta\theta' = -\frac{\eta}{2}\nabla L$
Original dynamics
$\Delta\theta' = -\frac{\eta}{2}\nabla F_\eta$
Effective dynamics
$\nabla F_\eta - \nabla L$
Entropic Force

Effective Free Energy

- Therefore, we have obtained an effective free energy:

$$F = L + \eta S$$

where

$$S = \left|\left|\nabla L\right|\right|^2$$

- Discrete-time training penalizes the gradient norm **(!!!)**

Stochastic Gradient Descent

$$S_{\mathrm{GD}} = |\nabla \mathrm{L}|^2$$
$$S_{\mathrm{SGD}} = \mathbb{E}[|\nabla \ell|^2]$$

Thus,

$$S_{\mathrm{SGD}} = S_{\mathrm{GD}} + \mathrm{Tr}[\mathrm{cov}(\nabla \ell, \nabla \ell)]$$

Entropy production
due to discretization

Entropy production due
to stochastic sampling

A closer look at the entropic term

For a fully connected layer:

$$p = Wh$$

$$S(W) = \mathbb{E}\left[\left|\nabla_p \ell\right|^2 |h|^2\right]$$

Simpler gradient

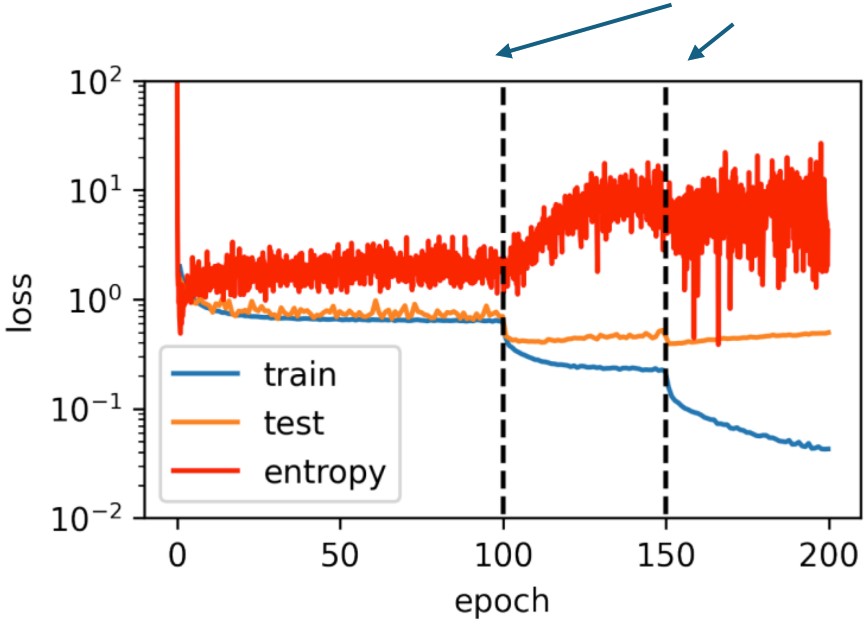
Simpler representation

Implication

- The learning dynamics at a small learning rate is qualitatively different that at a large learning rate
- A large learning rate regularizes the entropy (gradient norm)

Learning rate decrease

Resnet 18



Contents

1. Learning dynamics
2. Entropic Force Breaks Continuous Symmetries
3. Platonic Representation Hypothesis
4. Progressive Sharpening

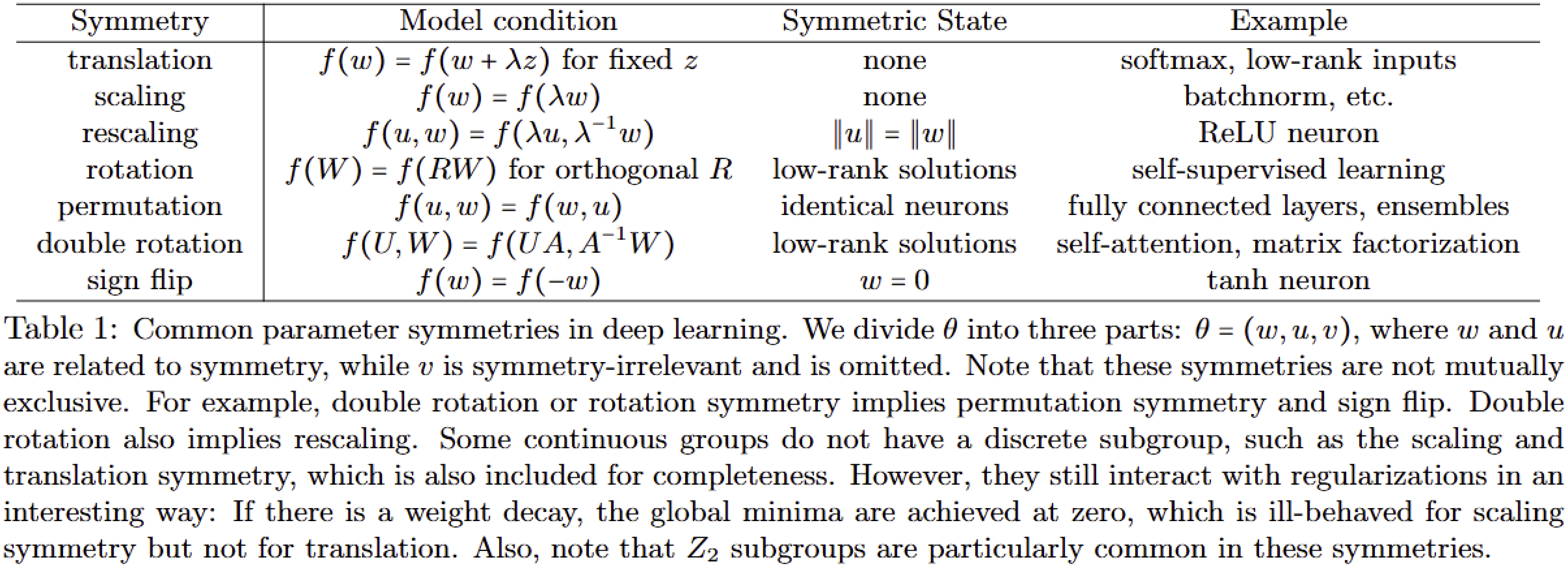
Parameter Symmetries

- Deep learning is full of parameter symmetries

Definition. Let G be a group. The loss function $L(\theta)$ has a G-symmetry if

$$L(\theta) = L(g\theta)$$

for all θ and $g \in G$.



Symmetry	Model condition	Symmetric State	Example
translation	$f(w) = f(w + \lambda z)$ for fixed z	none	softmax, low-rank inputs
scaling	$f(w) = f(\lambda w)$	none	batchnorm, etc.
rescaling	$f(u, w) = f(\lambda u, \lambda^{-1} w)$	$\lVert u \rVert = \lVert w \rVert$	ReLU neuron
rotation	$f(W) = f(RW)$ for orthogonal R	low-rank solutions	self-supervised learning
permutation	$f(u, w) = f(w, u)$	identical neurons	fully connected layers, ensembles
double rotation	$f(U, W) = f(UA, A^{-1} W)$	low-rank solutions	self-attention, matrix factorization
sign flip	$f(w) = f(-w)$	$w = 0$	tanh neuron

Table 1: Common parameter symmetries in deep learning. We divide θ into three parts: $\theta = (w, u, v)$, where w and u are related to symmetry, while v is symmetry-irrelevant and is omitted. Note that these symmetries are not mutually exclusive. For example, double rotation or rotation symmetry implies permutation symmetry and sign flip. Double rotation also implies rescaling. Some continuous groups do not have a discrete subgroup, such as the scaling and translation symmetry, which is also included for completeness. However, they still interact with regularizations in an interesting way: If there is a weight decay, the global minima are achieved at zero, which is ill-behaved for scaling symmetry but not for translation. Also, note that Z_2 subgroups are particularly common in these symmetries.

arxiv/2502.05300

Some examples

- Scaling symmetry:

$$f(W) = \frac{Wx}{||W||}$$

$$\rightarrow f(\lambda W) = f(W)$$

- This is also an example of a **non-compact Lie-group symmetry**
    - Rare in physics, but ubiquitous in deep learning

Parameter Symmetries

Theorem. F_η does not have any non-compact Lie-Group symmetry.

Theorem. Any symmetry of F_η must be norm-preserving.

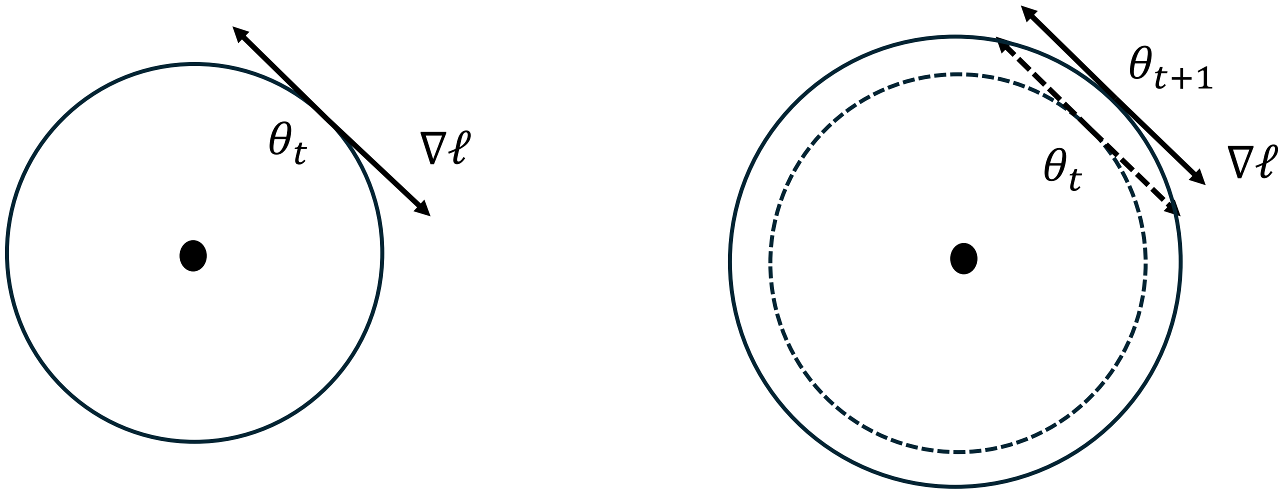
- Only rotation symmetries (discrete or continuous) will remain in F_η
    - Spontaneous symmetry breaking remains possible

Scale Invariance: An Example

- Consider a 2d problem with scale invariance: $\ell(\theta) = \ell(\lambda\theta)$
  - Conservation law under GF: $\frac{d}{dt}||\theta|| = 0$

Scale Invariance: An Example

- Consider a 2d problem with scale invariance: $\ell(\theta) = \ell(\lambda\theta)$
  - Conservation law: $\frac{d}{dt}||\theta|| = 0$
- The gradient $\nabla\ell$ must be tangent to conservation laws



Step t

Step $t+1$

Scale invariance →
A systematic flow towards infinity

Breaking of Rescaling Symmetry

- Consider a loss $\ell(\theta)$ with $\theta = (u, v)$
- Rescaling Symmetry:

$$\ell(\lambda u, \lambda^{-1} v) = \ell(u, v)$$

$$F = \underbrace{L}_{\text{Invariant}} + \underbrace{\eta\mathbb{E}||\nabla\ell||^2}_{\text{Variant}}$$

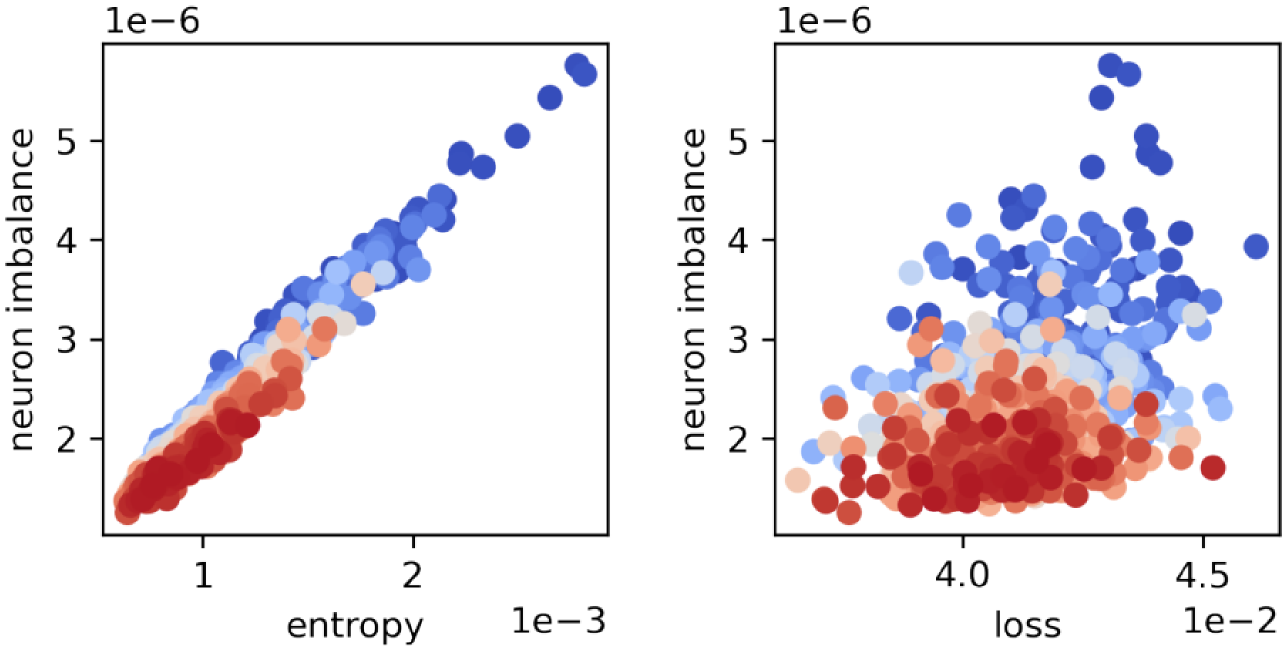
Theorem. All local minima of F satisfies **(neuron balance)**

$$||\nabla_{\mathrm{u}}\ell||^2 = ||\nabla_v\ell||^2$$

- An “equipartition” theorem: $S = ||\nabla_{\mathrm{u}}\ell||^2 + ||\nabla_v\ell||^2$

Example

- A ReLU network: $f(x) = \sum_i u_i \sigma\left(w_i^T x\right)$
  - $\sigma(z) = \max(0, z)$



Breaking of Generic Symmetry

- Exponential symmetry: for fixed symmetric matrix A, and any $\lambda \in \mathbb{R}$, $\ell(x,\theta) = \ell(x, e^{\lambda A}\theta)$

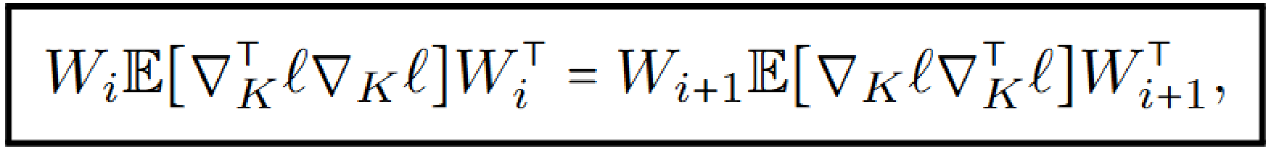
Theorem. If F_η has the A-exponential symmetry, all local minima of F_η satisfy

$$\mathbb{E}[\nabla_\theta^T \ell A \nabla_\theta \ell] = 0.$$

- Noise in different subspaces must balance!

Examples

- Matrix rescaling invariance, $\ell(W_{i+1}, W_i) = \ell(W_{i+1}B, B^{-1}W_i)$



$$W_i \mathbb{E}[\nabla_K^\top \ell \nabla_K \ell] W_i^\top = W_{i+1} \mathbb{E}[\nabla_K \ell \nabla_K^\top \ell] W_{i+1}^\top,$$

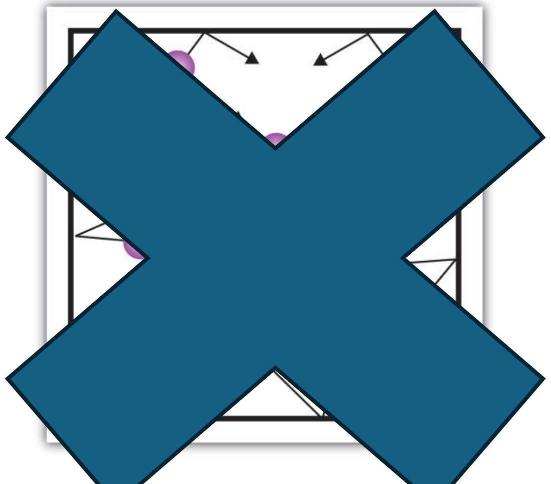
where $K = W_{i+1}W_i$

- e.g., $f(x) = UWx$

Implication

- The stationary distribution of SGD cannot be Gibbs
  - Not every state with the same energy (L) has the same probability of being accessed: **loss of ergodicity**
- **SGD is out-of-equilibrium**
  - Actually, SGD dynamics is almost everywhere *absolutely irreversible*

Murashita et al., PhysRevE.90.042110

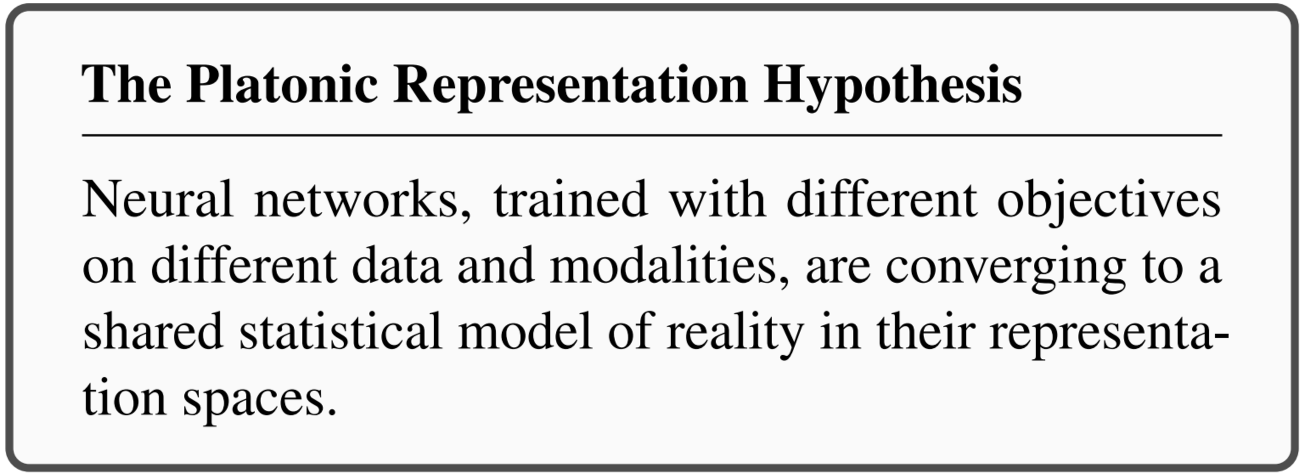


Contents

1. Learning dynamics
2. Entropic Force Breaks Continuous Symmetries
3. Platonic Representation Hypothesis
4. Progressive Sharpening

Platonic Representation Hypothesis

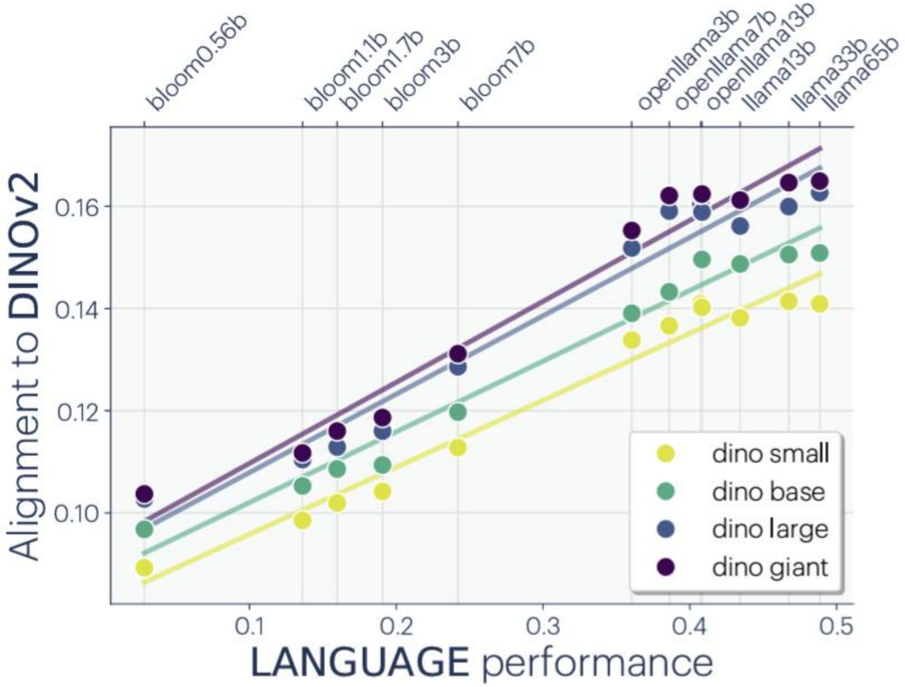
- Learned representations are universal (Platonic Representation Hypothesis)



The Platonic Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.

Huh et al., 2024



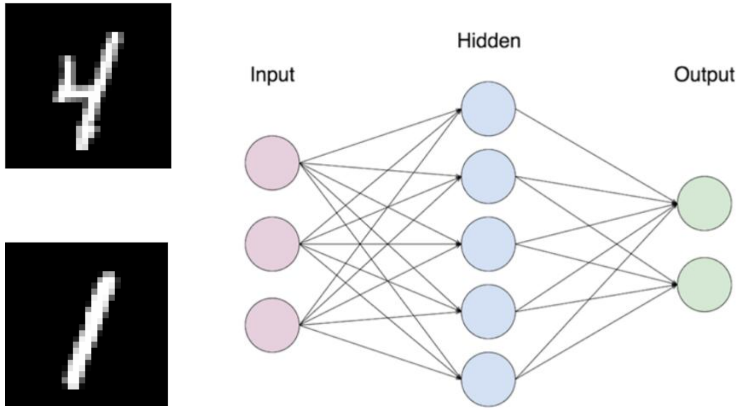
Perfect Platonic Representation Hypothesis

(NC is a special case)

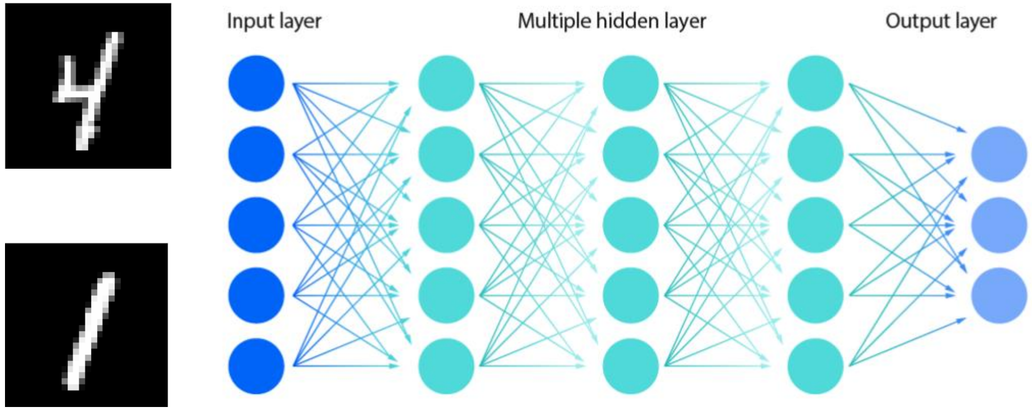
- We say that two layers h_A, h_B from two models A and B are perfectly aligned if

$$h_A^T(x_1)h_A(x_2) = h_B^T(x_1)h_B(x_2)$$

$\boldsymbol{h_A}$



$\boldsymbol{h_B}$



Universal Representation Learning in EDLN

- Consider a **Embedded Deep Linear Network (EDLN)** model (with a lot of symmetries)

$$f(x) = M^O W_D W_{D-1} \dots W_1 M^I x$$

- W: learnable, M: frozen invertible matrices
- Trained on different views of the same data:

$$D_A = \{(Z_A x_i, y_i)\}_i$$

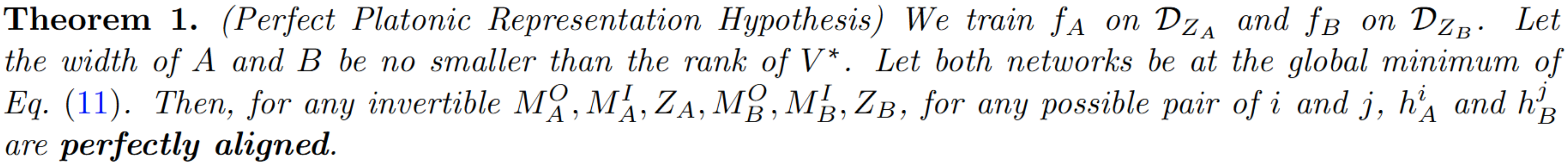
- Z_A: frozen invertible matrix
- MSE loss:

$$\ell(x, y, \theta) = \left|\left|f(x) - y\right|\right|^2$$

- $y = V^* x + \epsilon$

Theorem (informal). Train two different embedded deep linear net *A, B* that are wider than the target function. Give them different views of the same data, D_A, D_B. If the training reaches the global minimum of $\boldsymbol{F_\eta}$, then

1. *All layers of A are perfectly aligned with all layers of A*
2. *All layers of B are perfectly aligned with all layers of B*
3. *All possible pairs of layers between A and B are perfectly aligned*

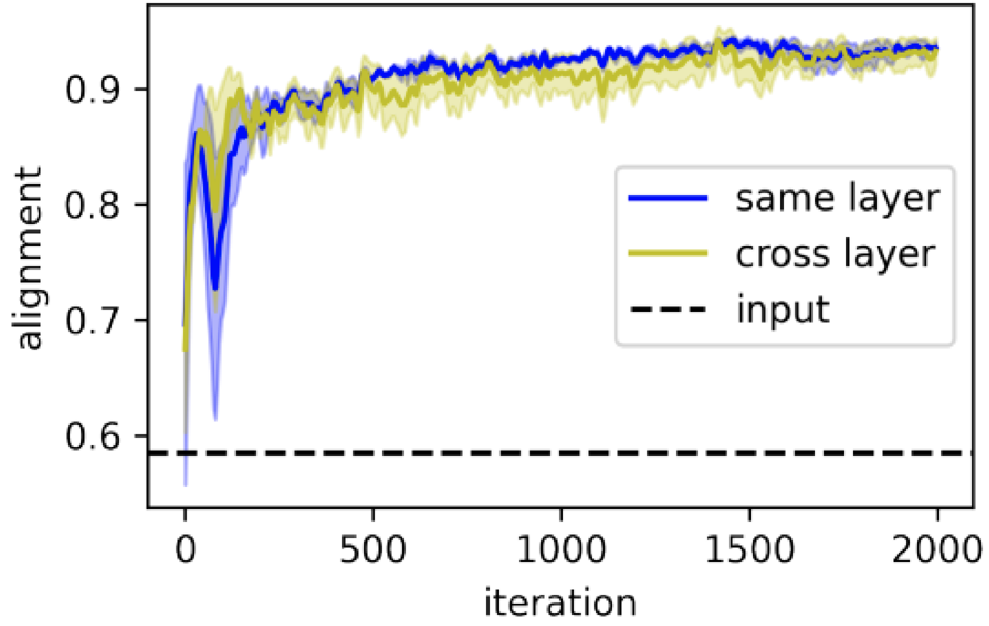


Theorem 1. *(Perfect Platonic Representation Hypothesis) We train f_A on $\mathcal{D}_{Z_A}$ and f_B on $\mathcal{D}_{Z_B}$. Let the width of A and B be no smaller than the rank of V^*. Let both networks be at the global minimum of Eq. (11). Then, for any invertible $M_A^O, M_A^I, Z_A, M_B^O, M_B^I, Z_B$, for any possible pair of i and j, h_A^i and h_B^j are* ***perfectly aligned****.*

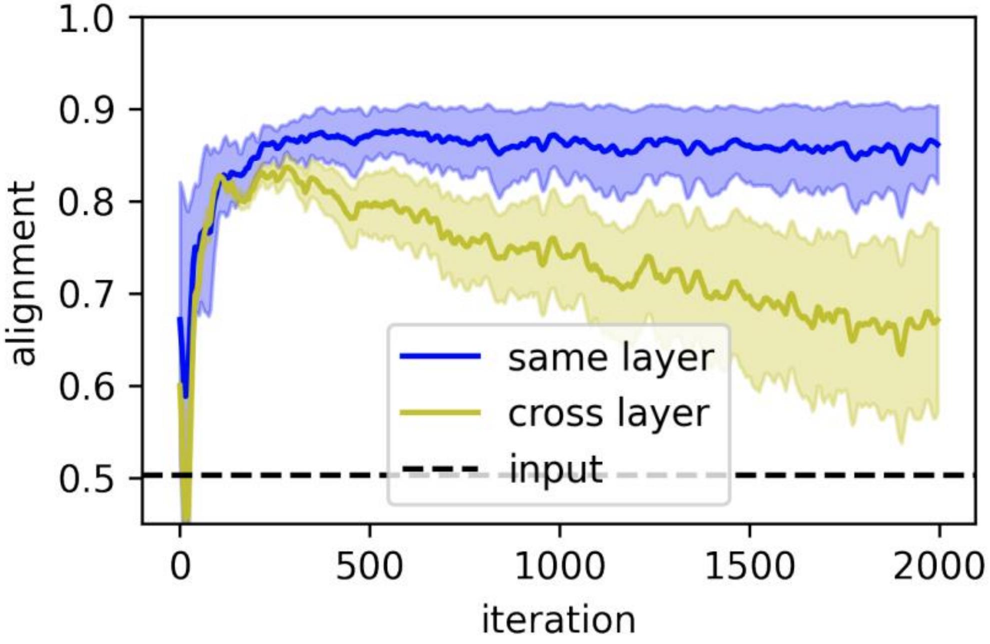
- Proof Sketch:

$$\boxed{W_i \mathbb{E}[\nabla_K^\top \ell \nabla_K \ell] W_i^\top = W_{i+1} \mathbb{E}[\nabla_K \ell \nabla_K^\top \ell] W_{i+1}^\top,}$$

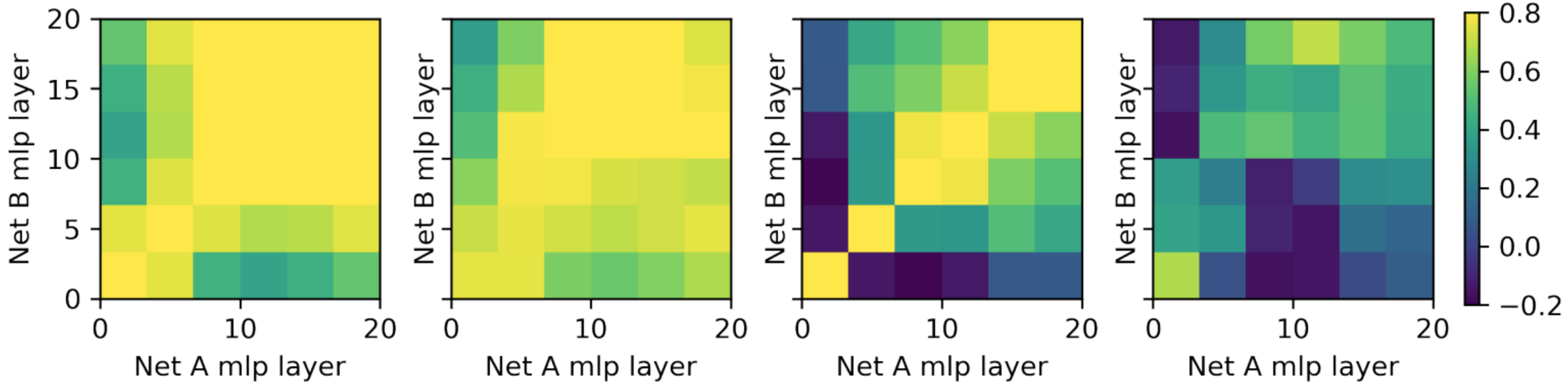
Deep linear networks



Nonlinear networks



Universal Representation Learning



Most solutions are not universal

- Let $\theta = (W_1, \dots, W_D)$ be a global minimum of L and one of its layer h_A is aligned with h_B of another network:
$$h_A^T(x_1)h_A(x_2) = h_B^T(x_1)h_B(x_2)$$
- Then, we can transform the layer before h_A by R and the layer after h_A by R^{-1}, and so the l.h.s. becomes
$$h_A^T(x_1)RR^T h_A(x_2)$$

which cannot be perfectly aligned to the r.h.s.

Universal Representation Learning

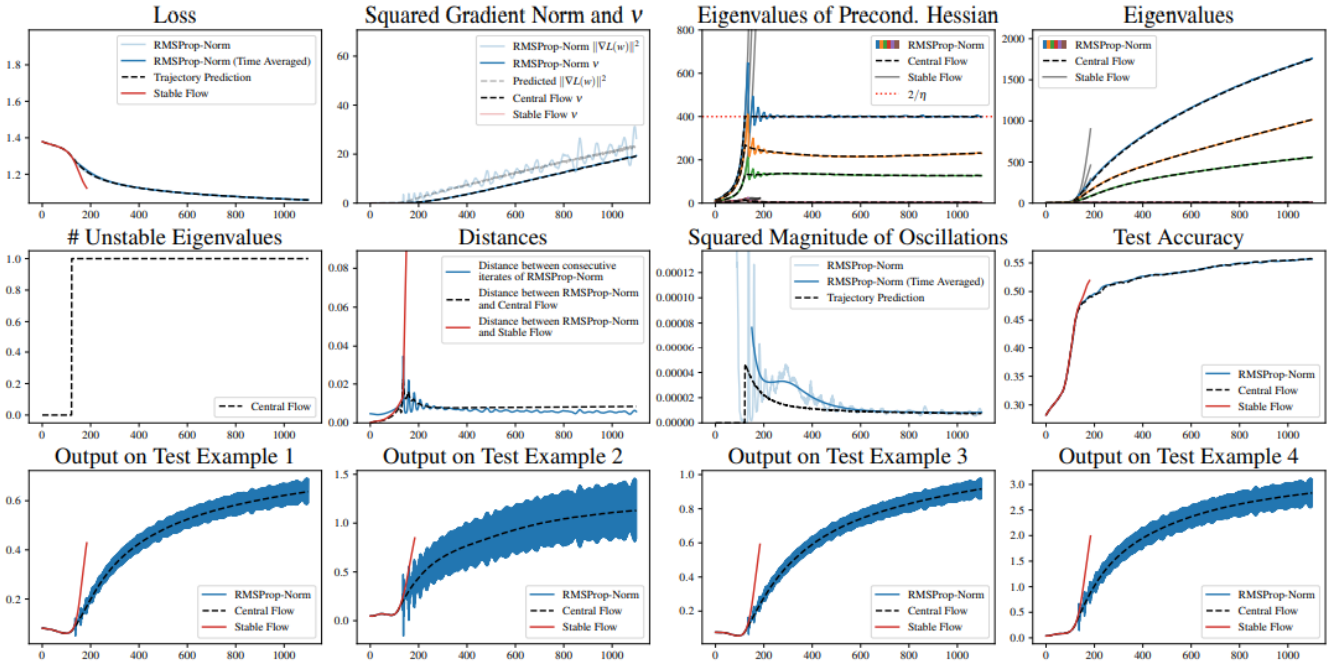
- This is extraordinary
  - There are infinitely many solutions (due to symmetry) that are not universal
- The mechanism is different from any of the previously conjectured ones (Huh et al., 2024):
  - Simplicity bias
  - Multitask training
  - Large capacity
- Symmetry and entropy are the cause
  - (Goldstone modes have a preferred orientation when there is non-compact Lie-group symmetries in the loss function)

Contents

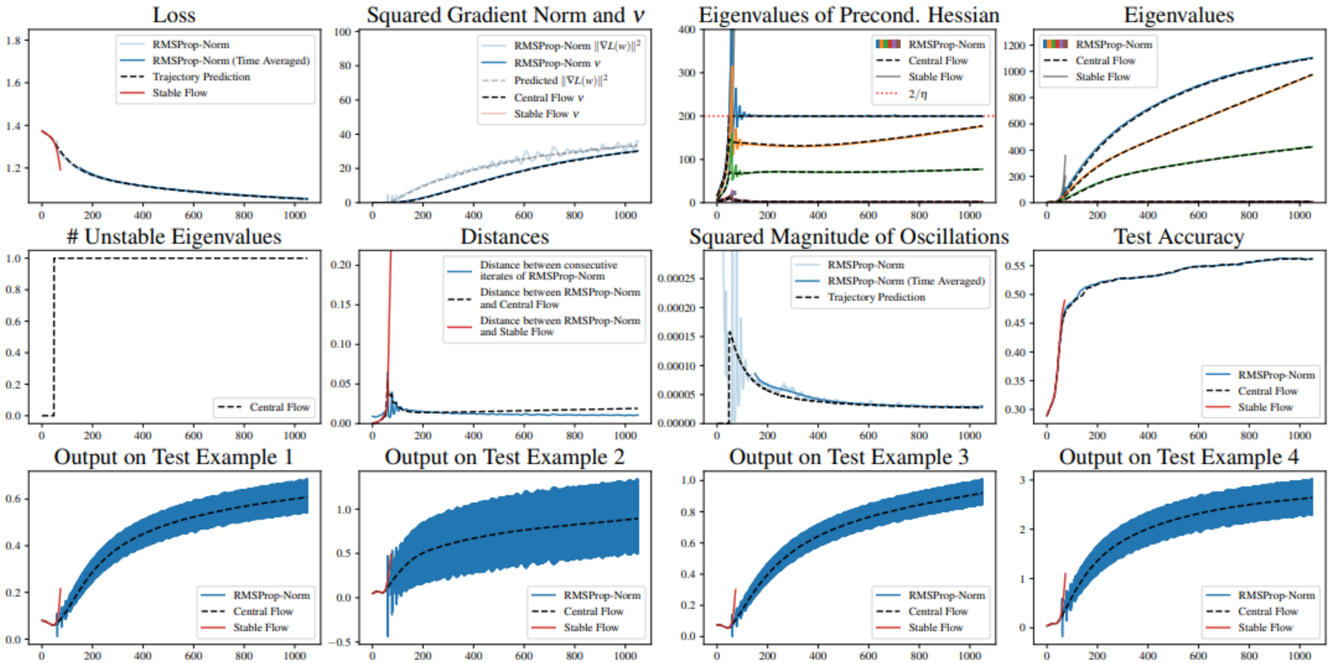
1. Learning dynamics
2. Entropic Force Breaks Continuous Symmetries
3. Platonic Representation Hypothesis
4. Progressive Sharpening

Progressive Sharpening (PS)

- A universal phenomenon
- The Hessian eigenvalues of the loss increases as the training proceeds
    - The learning dynamics gradually loses stability
    - Sharpness: the largest eigenvalue of the Hessian



(a) $\eta = 0.005$, $\beta_2 = 0.995$, $\epsilon = 10^{-7}$, bias correction



Cohen et al.,
arxiv/2410.24206

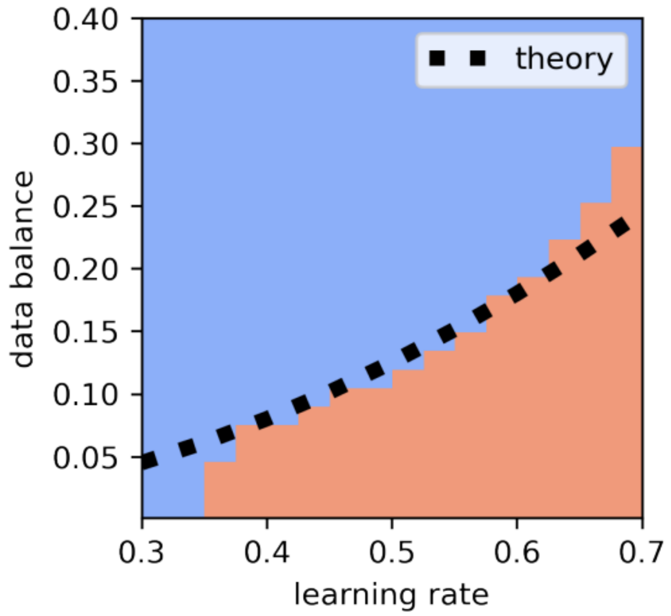
PS and PRH have the same cause

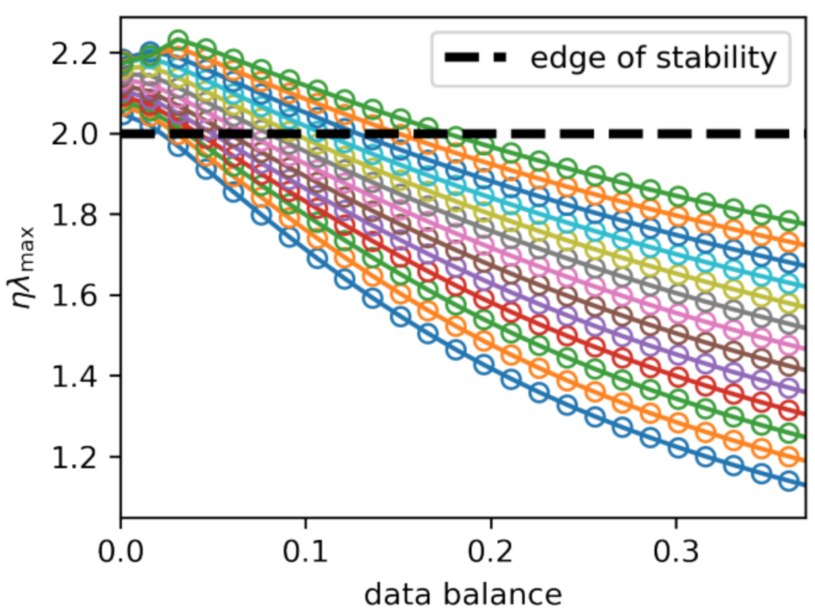
A zeroth-order intuition:

- In EDLN, the sharpness depends on the data distribution
- The learned solution is independent of data distribution
- There must exist data distribution for which the model converges to an arbitrarily sharp solution

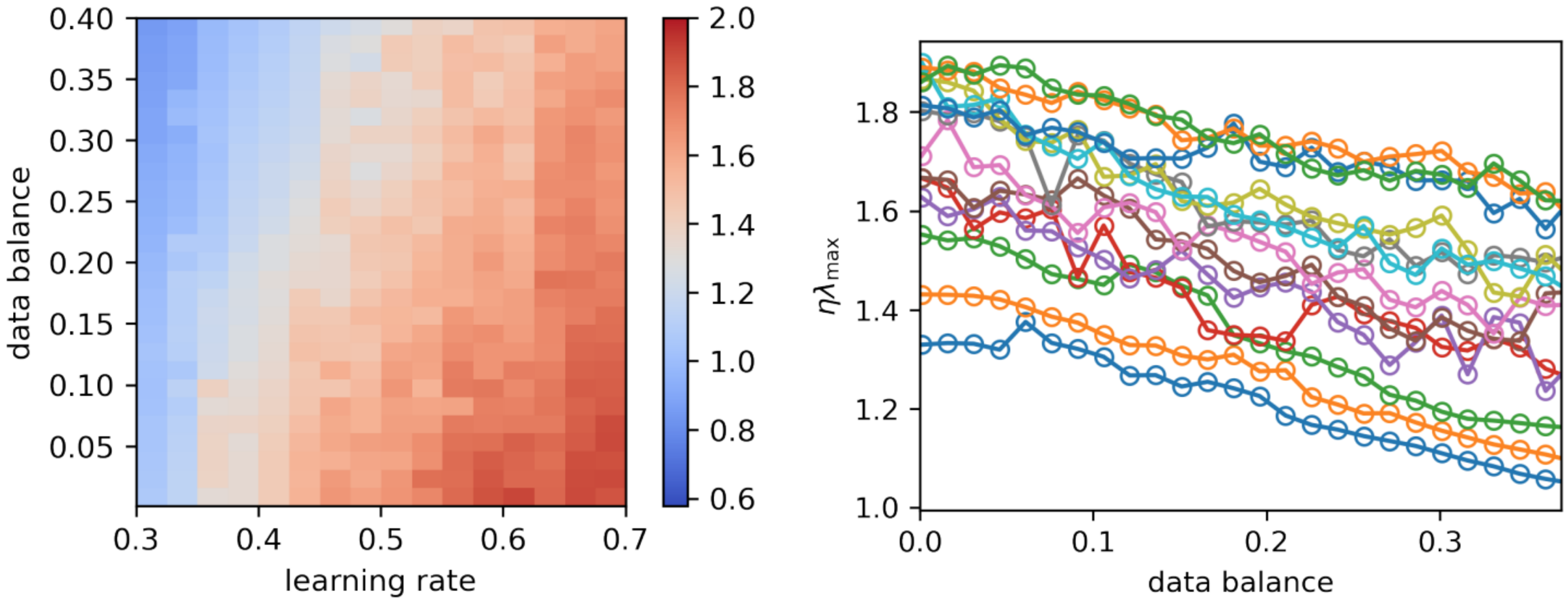
Prediction. Imbalance of label uncertainty leads to progressive sharpening.

- Balance of label uncertainty leads to progressive flattening
- In language, the uncertainty of next words is highly imbalanced





Nonlinear models



Summary

- The learning dynamics of modern AI models is irreversible
- This irreversibility gives rise to an **entropic force**, which plays a crucial role in representation formation
- The entropic force shows that two universal phenomena in representation learning have the same cause
    - PRH
    - Progressive sharpening
- Can we have a “**unification**” of theories of deep learning ???

A slightly broader picture

- A loss function generally takes the following form:

$$L = L_0 + \mathbf{Regularization} + \mathbf{Entropy}$$

Symmetries:
Structures,
Universality

Simpler weights

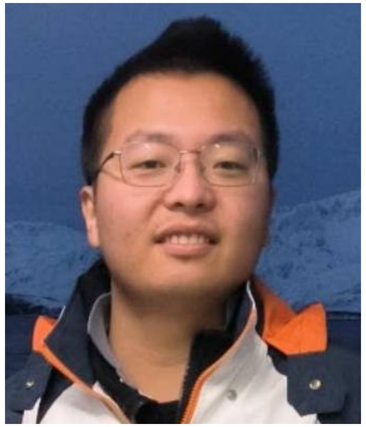
Simpler gradient
Simpler representation

A wild conjecture

Symmetry + Noise + Regularization ≈ **Deep Learning**

?

Thanks



Yizhou Xu

Isaac Chuang

Liu Ziyin*, Yizhou Xu*, Isaac Chuang. *Neural Thermodynamics I: Entropic Forces in Deep and Universal Representation Learning*. arxiv 2505.12387

- Proof sketch: Let $\theta \to e^{\lambda A}\theta$. L is invariant to this transformation, but $\nabla \ell$ transforms by
$$e^{-\lambda A}\nabla \ell.$$
- So,
$$\mathbb{E}[\nabla_\theta^T \ell \nabla_\theta \ell] \to \mathbb{E}[\nabla_\theta^T \ell e^{-2\lambda A} \nabla_\theta \ell].$$
- Decompose in two subspaces of A: $A = A_+ + A_-$,
$$\mathbb{E}[\nabla_\theta^T \ell e^{-2\lambda A_+} \nabla_\theta \ell] + \mathbb{E}[\nabla_\theta^T \ell e^{-2\lambda A_-} \nabla_\theta \ell].$$
- This term is minimized at a unique λ^* for every such A-symmetry.

Examples

- Scaling invariance: $A = I$,

$$\mathbb{E}[\nabla_\theta^T \ell \nabla_\theta \ell] = 0.$$

- Rescaling invariance: $A = (I_u, -I_v)$,

$$\mathbb{E}|\nabla_u \ell|^2 - \mathbb{E}|\nabla_v \ell|^2 = 0$$

Two-layer linear network can be exactly solved

$$\ell(\theta, x) = |W_2 W_1 x - y|^2$$

- $y = V^* x + \epsilon$