

How can physics help understand deep learning?

Liu Ziyin

MIT

NTT Research

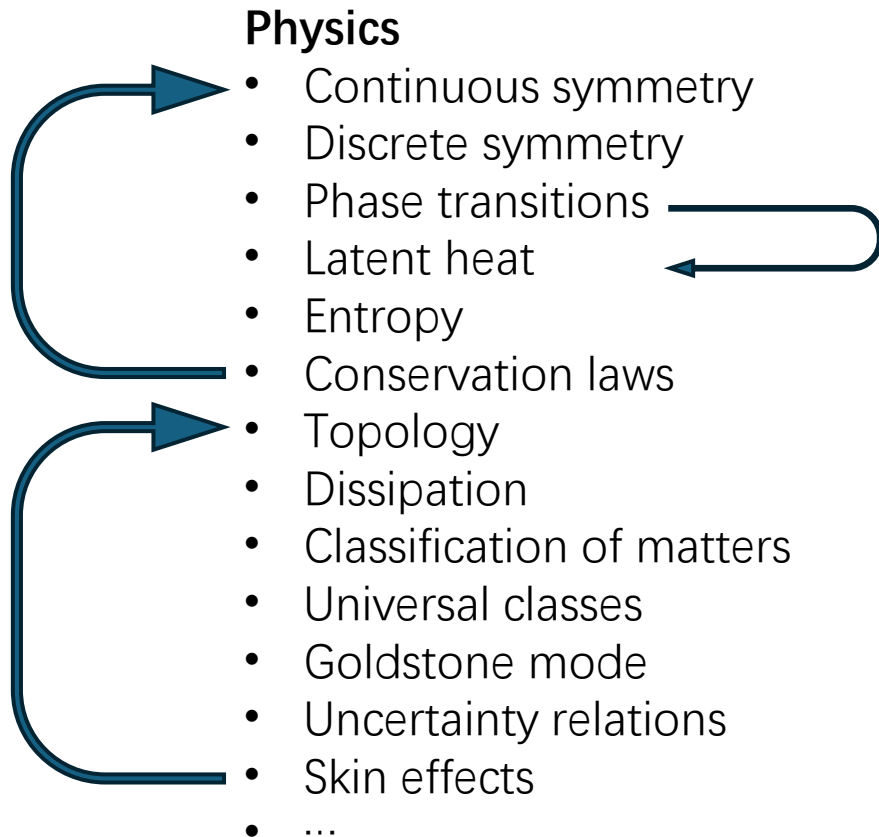
10/24/2024

Theories of physics?

- Newton's Law: $F = \frac{m_1 m_2}{r^2}$
- Special Relativity: $E = mc^2$
- Special Relativity (Part II): **Space** is (essentially) the same as **time**
- Noether's theorem: every **continuous symmetry** leads to a **conservation law**
- Landau Theory: **Phase transitions** are due to **change of symmetries**
- Fluctuation-Dissipation Theorem: **dissipation** must balance with **fluctuation**
- ...

Theories of physics?

- Theoretical physics involves creating simple concepts and connect them



Theories of AI?

- Weight decay improves [generalization](#) (Krogh & Hertz. 1991)
 - Weight decay determines the effective [learning rate](#) (arxiv/2010.02916)
 - Regularization is necessary for [neural collapse](#) (arxiv/2410.04887)
 - Overparametrized models can memorize all data and still [generalize](#) (arxiv/1611.03530)
 - All [local minima](#) are [global](#) (arxiv/1605.07110)
 - All [global minima](#) are connected in [overparametrized](#) networks (arxiv/1901.07417)
 - ...
-
- We also need to invent and [connect](#) concepts

Physics of AI?

- Link fundamental concepts of physics to those of AI

Physics

- Continuous symmetry
- Discrete symmetry
- Phase transitions
- Latent heat
- Entropy
- Conservation laws
- Topology
- Dissipation
- Classification of matters
- Universal classes
- Goldstone mode
- Uncertainty relations
- Skin effects
- ...

AI

- Reasoning
- Generalization
- Optimization
- Learning dynamics
- Overparametrization
- Scaling laws
- Neural collapse
- Neural feature ansatz
- Neural tangent kernel
- Feature learning
- Mode connectivity
- Emergence (of capabilities)
- ...

?



Fundamental Concepts of Physics?

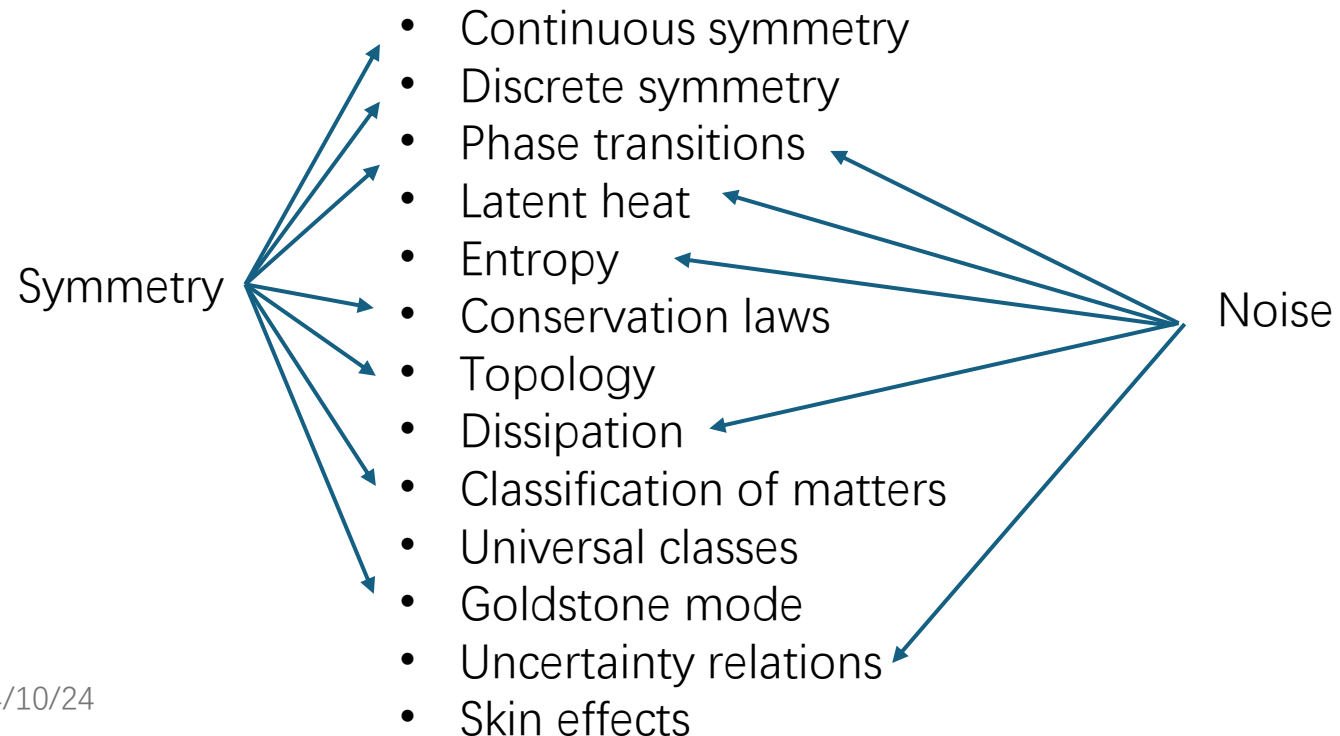
“It is only slightly overstating the case to say that physics is the study of **symmetry**. ” -- Philip W. Anderson

- **Noise** is perhaps the second most important concept

Fundamental Concepts of Physics?

“It is only slightly overstating the case to say that physics is the study of **symmetry**. ”
-- Philip W. Anderson

- **Noise** is perhaps the second most important concept:



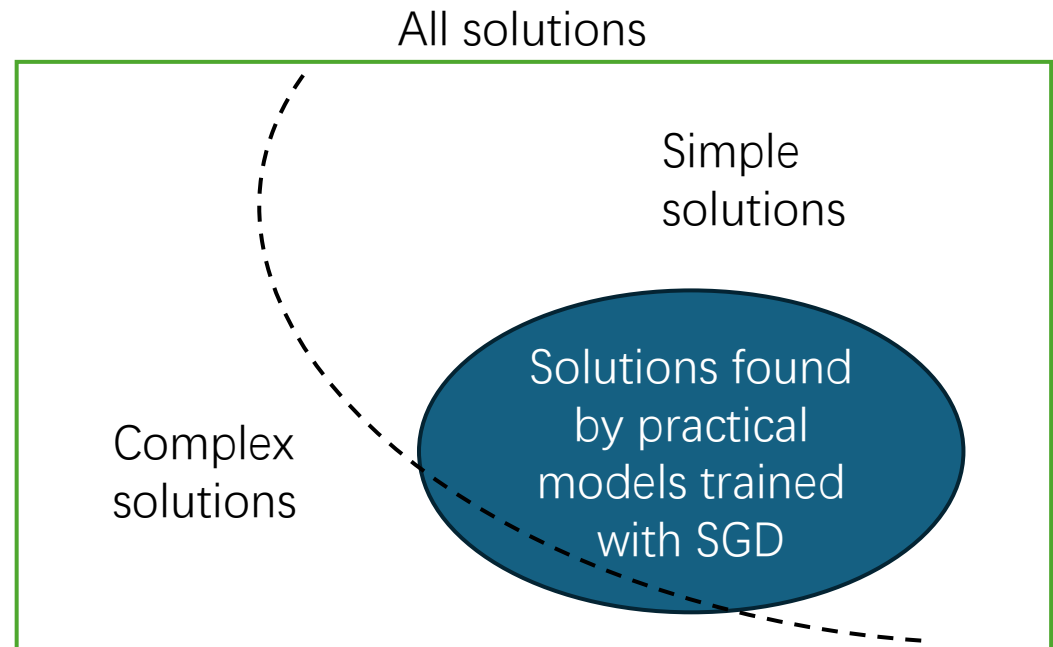
Fundamental Concepts of deep learning?

- If fully unconstrained, we have curse of dimensionality
 - Overparametrized models often find good enough solutions
- There are **simplicity biases** in deep learning

Fundamental Concepts of deep learning?

- If fully unconstrained, we have curse of dimensionality
 - Overparametrized models often find good enough solutions
- There are **simplicity biases** in deep learning

Why?



Hypothesis

Symmetry + Noise + Regularization \approx **Simplicity Bias**
(for Deep Learning)

Simplicity Triangle

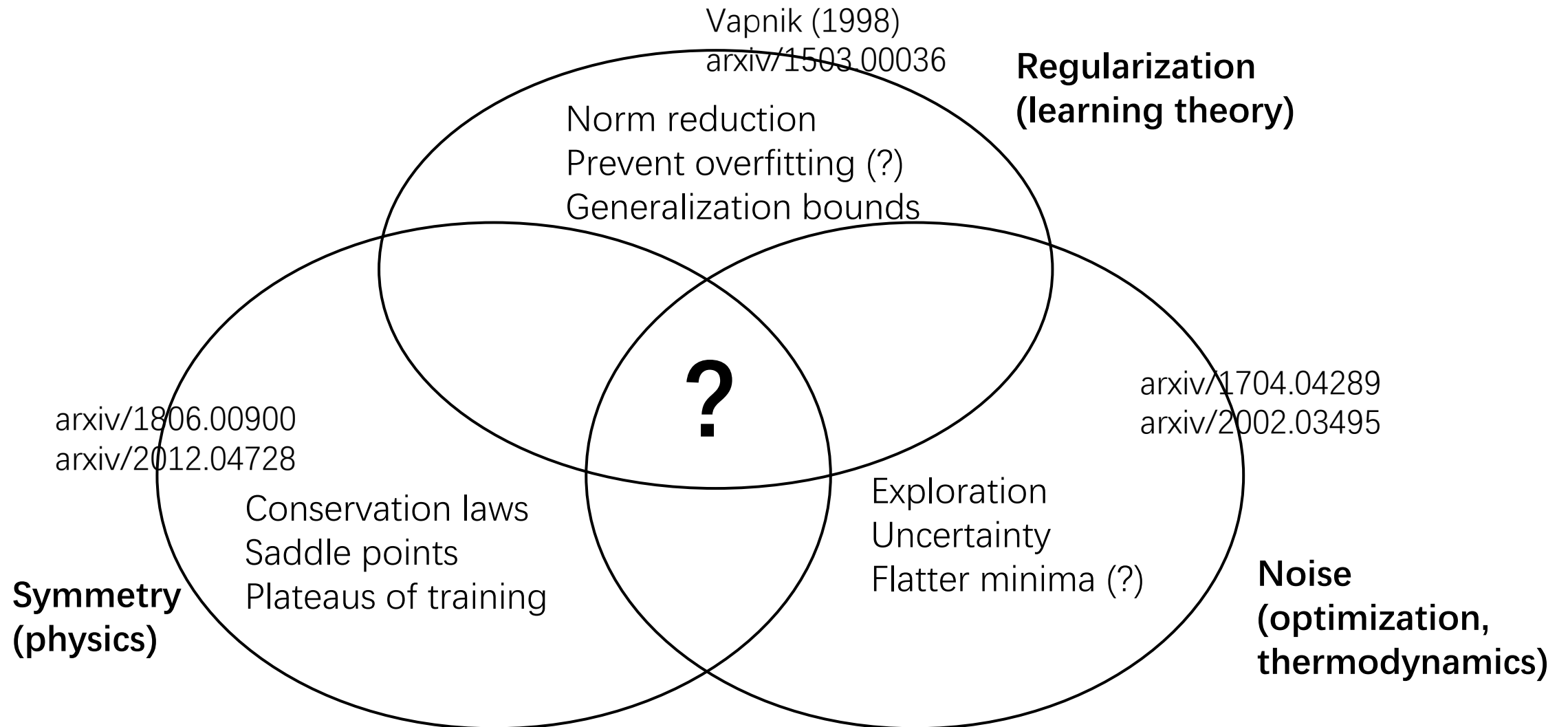


Table of contents

1. Physics of deep learning?
2. Discrete symmetry + Regularization = Constrained Parameters
3. Continuous Symmetry + Noise = Init. Independent Solutions
4. Noise + Regularization = Compact Representations

[1] Symmetry Induces Structure and Constraint of Learning.
ICML 2024

Two Types of symmetries in Deep Learning

1. Data symmetry:

- Equivariant networks

2. Parameter symmetry

- Can be leveraged to understand the learning dynamics and loss landscape of neural networks

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

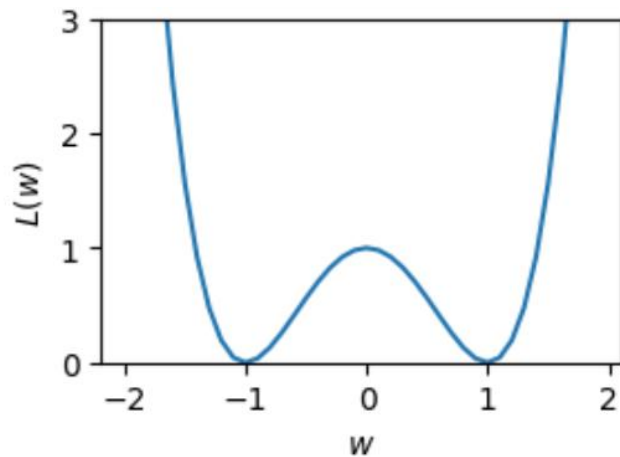
$$L(\theta) = L(g\theta)$$

for all θ and $g \in G$.

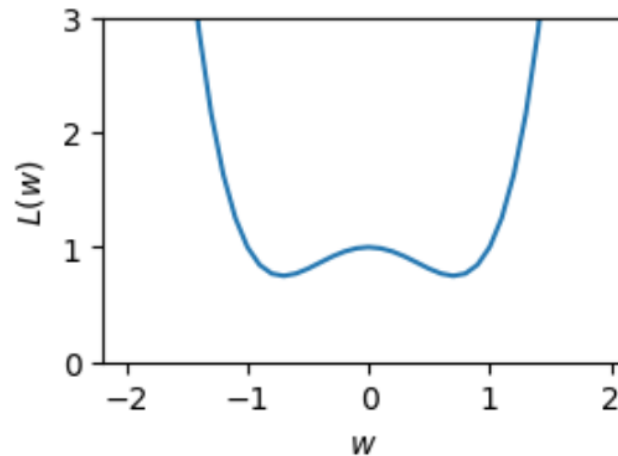
Landscape of models with symmetry

- Consider quadratic regression with L_2 regularization (weight decay):

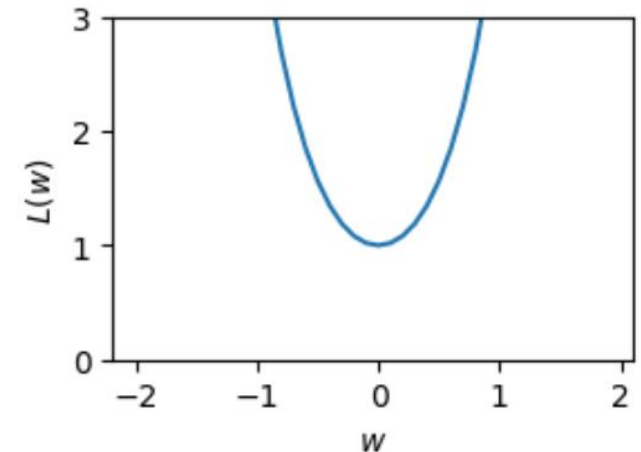
$$L(w) = (w^2x - y)^2 + \gamma w^2$$



Small γ



Intermediate γ



Large γ

Landscape of models with symmetry

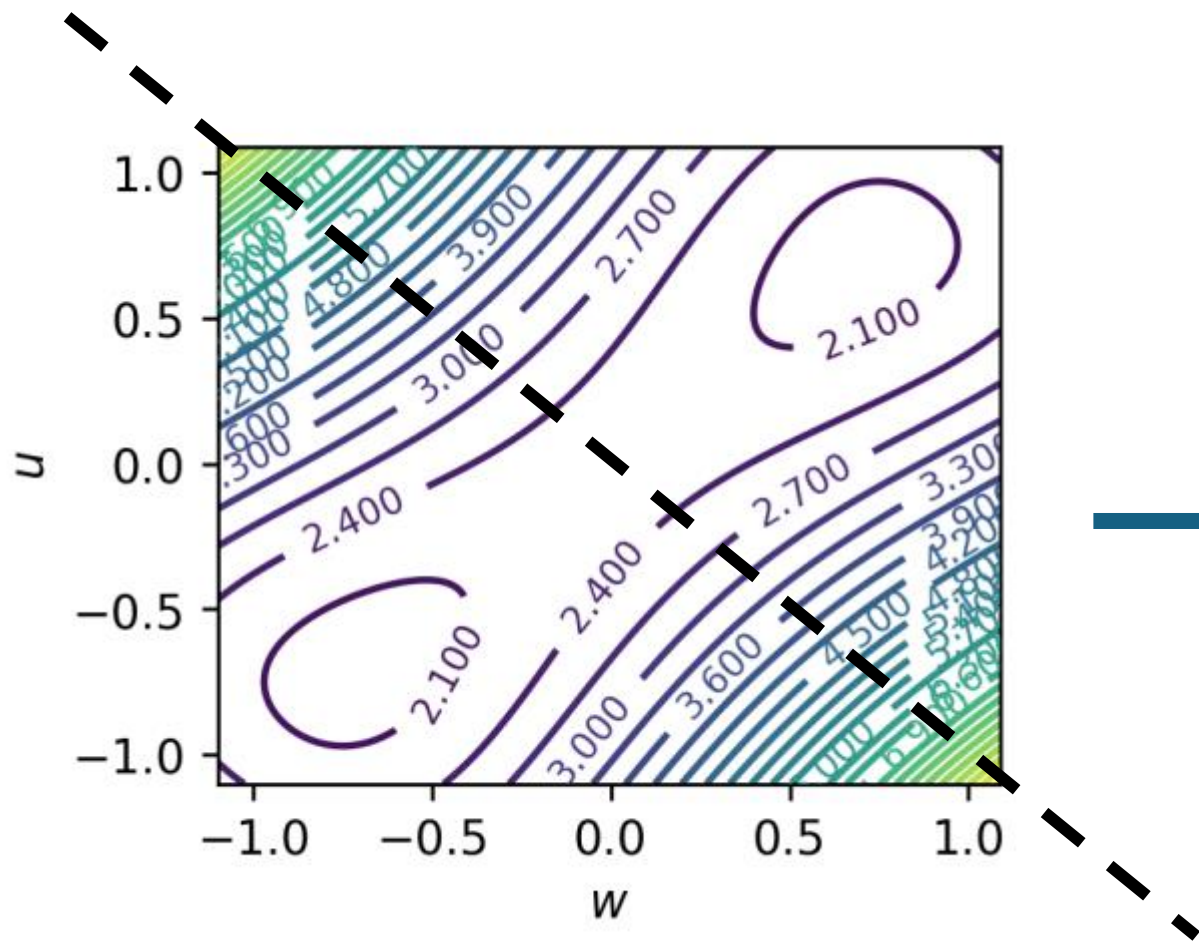
- The **symmetric solution** becomes the global minimum at **strong regularization**
 - Can be generalized to high dimension
 - Can be generalized to an arbitrary discrete group
- Setting:

$$L(\theta) = L_0(\theta) + \gamma ||\theta||^2$$

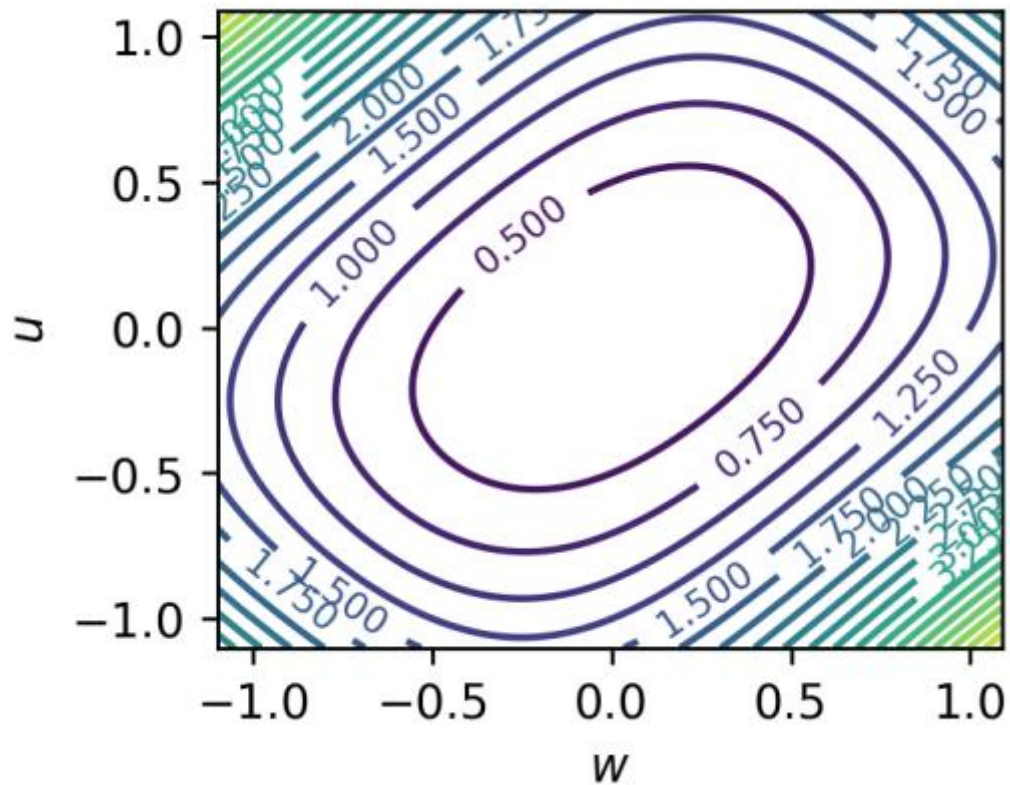
Theorem. (Informal) Let $L_0(\theta)$ have the G -symmetry. Then, if γ is large enough, all global minima θ^* of the $L(\theta)$ satisfy

$$g\theta^* = \theta^*$$

for any $g \in G$.



small γ



large γ

Abundance of Mirror symmetries

Symmetry	Loss	Symmetry Projector	Symmetric State
Rescaling invariance	$\ell_0(u, w) = \ell_0(\lambda u, \lambda^{-1} w)$	$OO^T = \begin{pmatrix} I_w & 0 \\ 0 & I_u \end{pmatrix}$	$u = 0, w = 0$
Rotation invariance	$\ell_0(W) = \ell_0(RW)$ for arbitrary orthogonal R	$OO^T =$ <i>arbitrary projection</i>	$n^T W = 0$ for arbitrary n
Permutation invariance	$\ell_0(u, w) = \ell_0(w, u)$	$OO^T = \begin{pmatrix} 0 & I_u \\ I_w & 0 \end{pmatrix}$	$w = u$

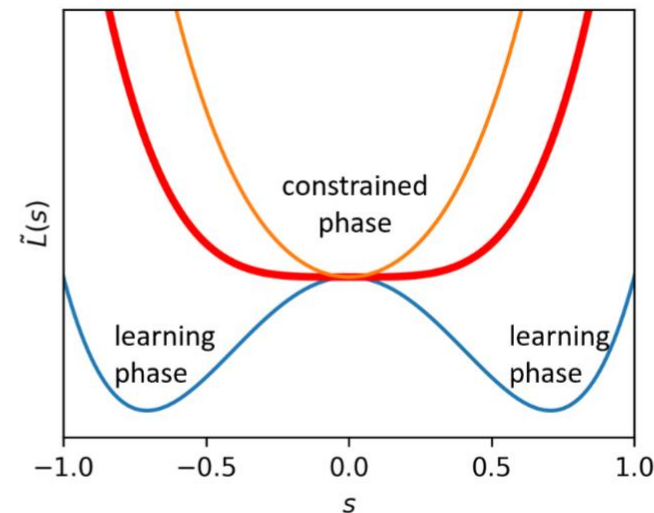
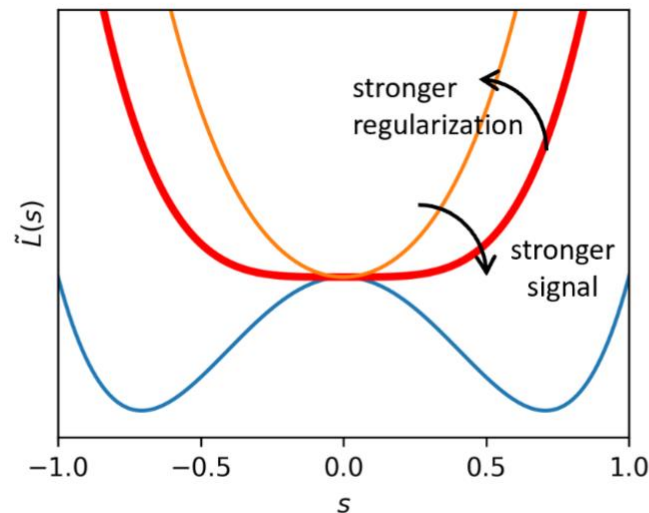
- In words,
 - Rescaling symmetry → sparsity
 - Rotation symmetry → low rankness
 - Permutation symmetry → identical neurons
- With L_2 regularization, every discrete symmetry leads to a structured constraint of learning

Implication for the loss landscape

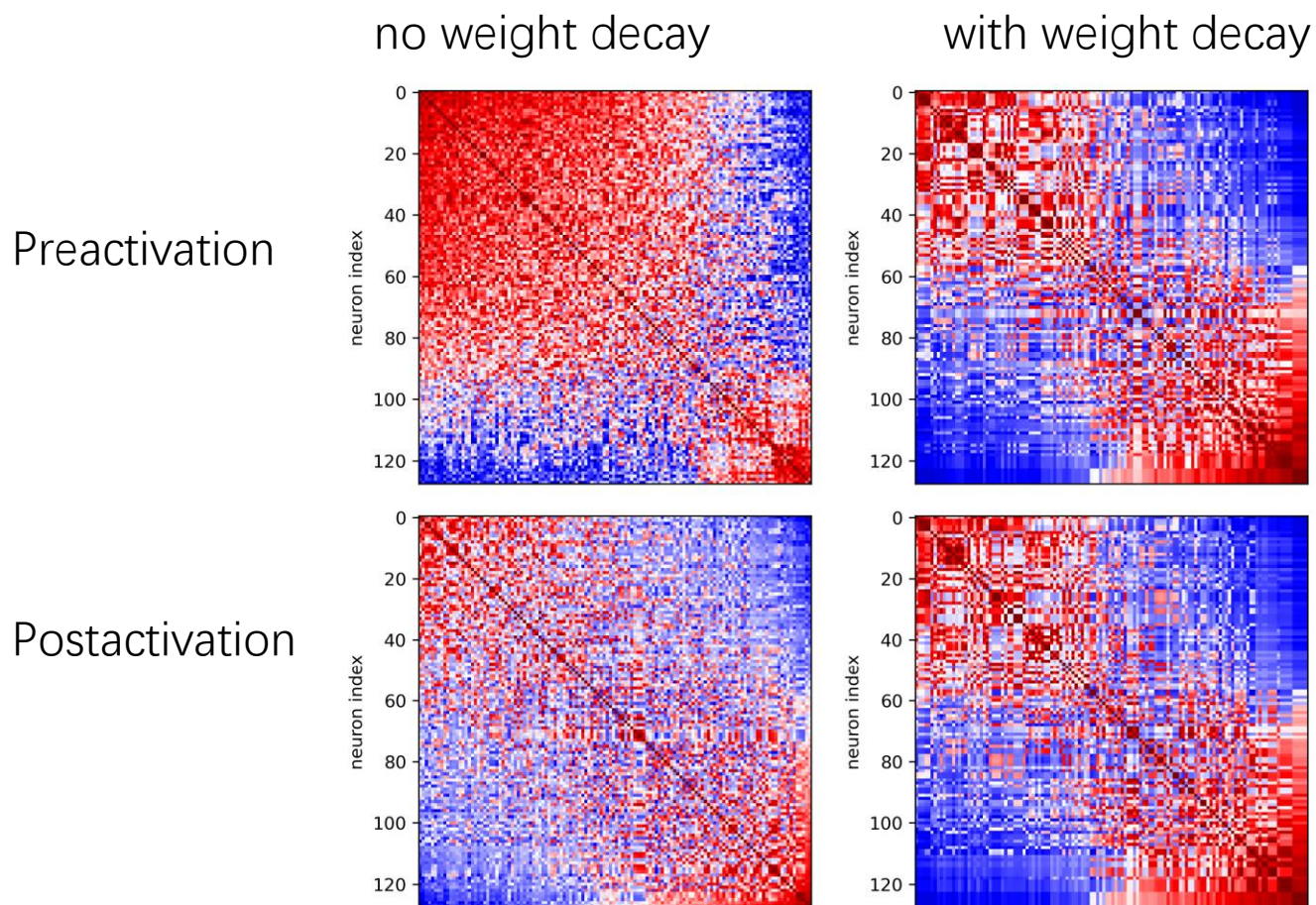
- When \mathbb{Z}_2 symmetries exist, 1d projections of the loss landscape are symmetric around the **mirror surface** (let $\mathbf{O}^T \mathbf{w} = sn$, for a unit vector n)

$$\ell(s) = c_0 s^2 + c_1 s^4 + O(s^6)$$

- Sign of c_0 determines the local geometry



Activation pattern of ResNet18



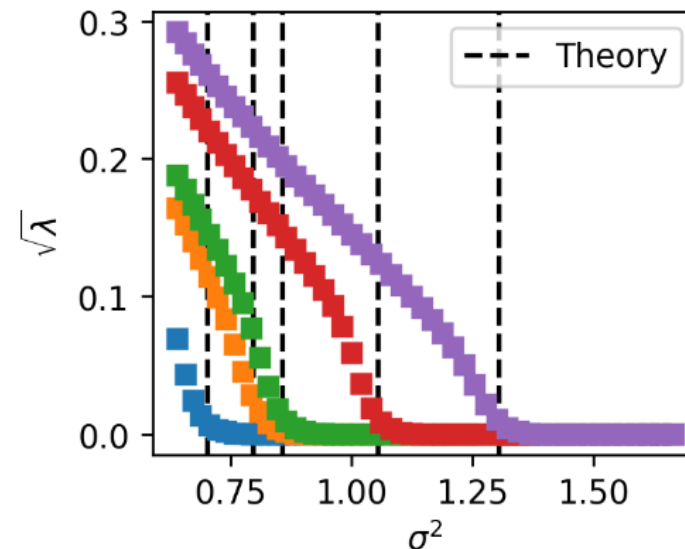
Preactivation and postactivation have a similar rank

Figure 5: Comparison for the correlation matrix of the neurons in the penultimate layer at zero weight decay (**left**) and 0.001 weight decay (**right**). **Upper**: pre-activation correlation. **Lower**: post-activation correlation. After training, the neurons are grouped into homogeneous blocks when weight decay is present. The inset shows that such block structures are very rare when there is no weight decay. Also, the patterns are similar for post-activation values, which further supports the claim that the block structures are due to the symmetry, not because of linearity.

Dimensional Collapse in Self-Supervised Learning

- The SimCLR loss has a rotation symmetry between the data point pairs

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}, \quad (1)$$



*Ziyin et al. *What shapes the loss landscape of self-supervised learning?*
ICLR 2023

Posterior collapse in Bayesian deep learning

- The ELBO loss is invariant to a simultaneous rotation of the decoder input and incoder output

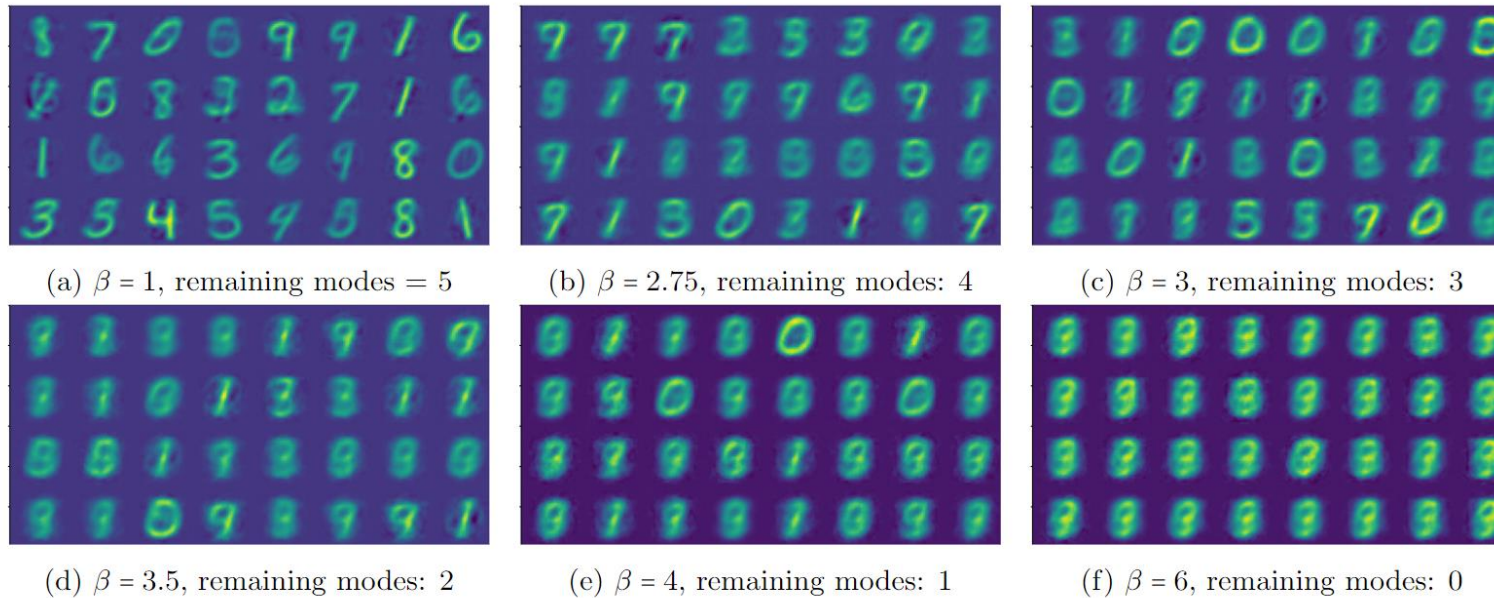
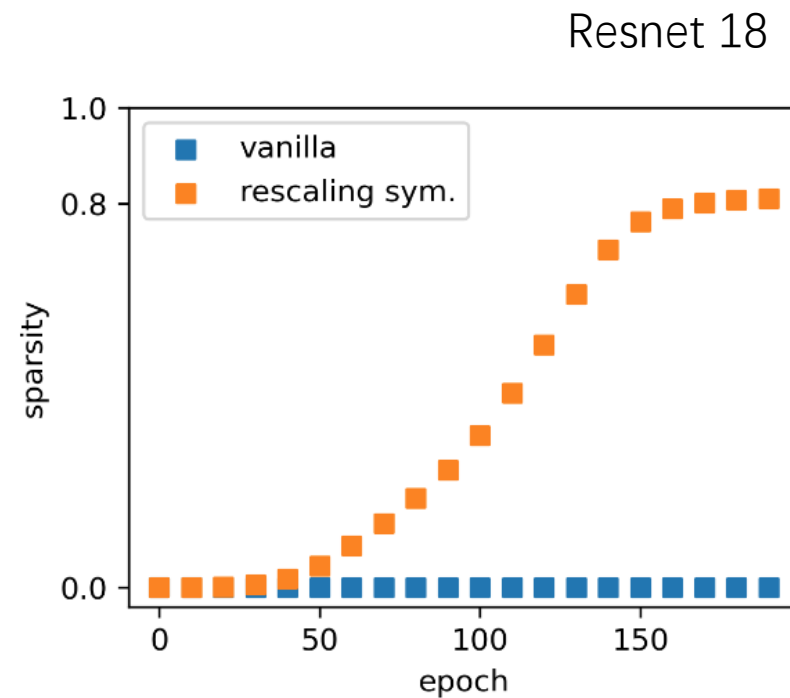
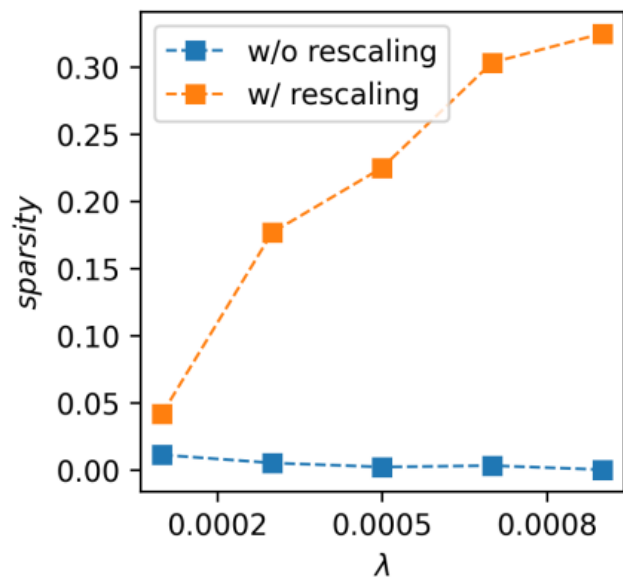


Figure 3: MNIST generation under different β . We see that the generated images lose diversity and variation as β increases. The number of mode left is estimated by the theoretical prediction of thresholds of each singular values.

*Wang et Ziyin. *Posterior collapse in a latent variable model*. NeurIPS 2022

Removing symmetry also Removes the Constraint



Message

- Discrete symmetries creates **constraints** over model parameters and model capacity
- In physics terms, collapses are transitions from symmetry-broken states to symmetry states
 - One can leverage symmetry to design training algorithms (arxiv/2210.01212)

[1] Symmetry Induces Structure and Constraint of Learning.
ICML 2024

Table of contents

1. Physics of deep learning?
2. Discrete symmetry + Regularization = Constrained Parameters
3. Continuous Symmetry + Noise = Init. Independent Solutions
4. Noise + Regularization = Compact Representations

[2] Loss Symmetry and Noise Equilibrium of Stochastic Gradient Descent. *NeurIPS 2024*

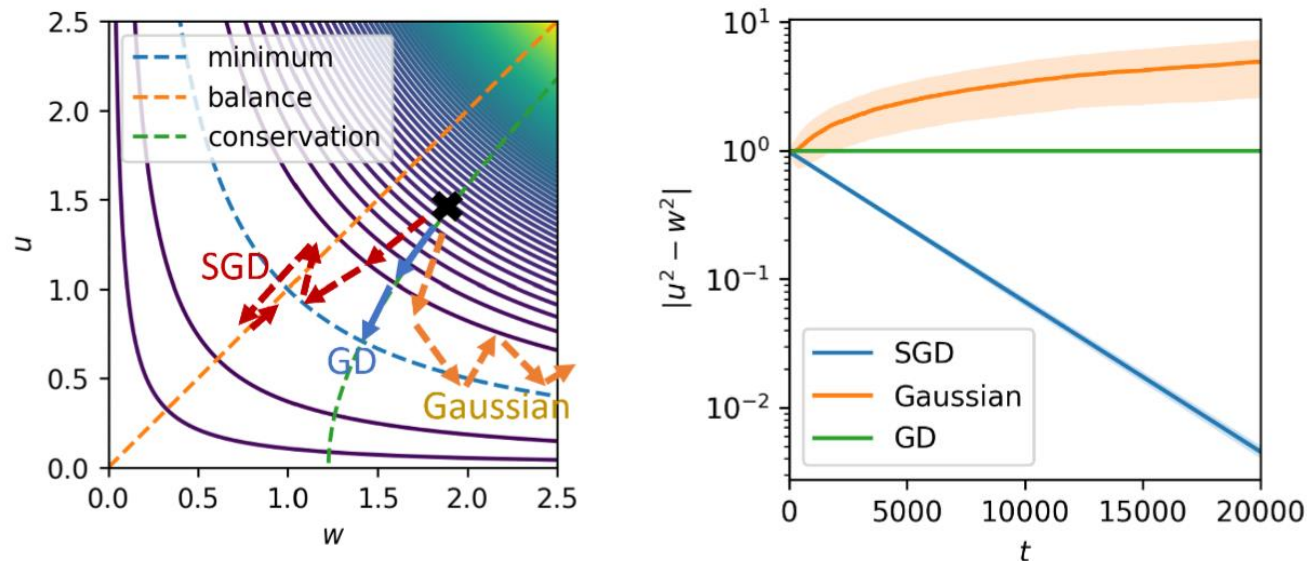
An example

- Consider a simple example with rescaling symmetry:

$$\ell(u, w) = (uwx - 1)^2$$

where $x \sim N(0,1)$.

- Under gradient flow, $u^2 - w^2$ does not change in training



Lagrangian Formalism

- Training proceeds with continuous-time gradient descent,

$$\dot{\theta} = -\nabla_{\theta} L(\theta)$$

- The training loss itself is the Lagrangian of the system (Bregman Lagrangian):

$$\mathcal{L} = L(\theta)$$

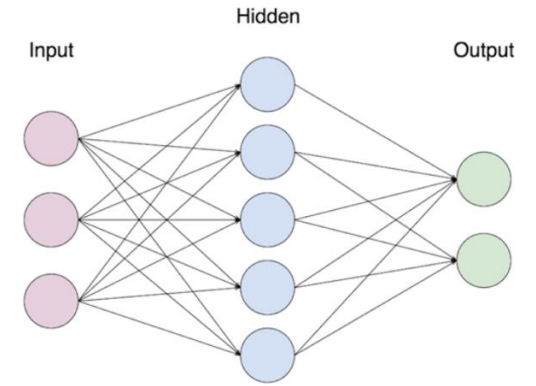
Noether's Theorem. (Exaggerated) Every continuous symmetry of the loss ($L(\theta) = L(g(\theta))$) leads to a conserved quantity \mathcal{C}

$$\frac{d}{dt} \mathcal{C}(\theta) = 0$$

- Under GD, the learned solution is dependent on the initialization

Two common continuous symmetry

- Continuous Symmetries in deep learning
 - **Rescaling symmetry**: $L_0(u, w) = L_0(\lambda u, \lambda^{-1} w)$
 - ReLU activation, linear models
 - The norm difference is conserved: $|u|^2 - |w|^2$
 - **Double rotation symmetry**: $L_0(U, W) = L_0(UA^{-1}, AW)$
 - Matrix factorization ($f(x) = UWx$), transformers
 - The matrix product is conserved: $U^T U - WW^T$



Breakdown of Noether's Theorem

- In actual model training, the training process is always **stochastic**
- Even if $\ell(\theta)$ has a symmetry with probability 1, Noether's theorem is no longer applicable
symmetry \neq conserved charge
- Can we still say something universal?

General Theory

- Let $\ell(\theta)$ have an **exponential symmetry**: with probability 1, for a fixed symmetric matrix A and any $\lambda \in \mathbb{R}$,
$$\ell(\theta) = \ell(e^{\lambda A} \theta)$$

General Theory

- Noether Charge: $C_A(\theta) = \theta^T A \theta$
 - Under continuous-time GD: $\frac{d}{dt} C_A = 0$
 - Under SGD: $\frac{d}{dt} C_A = \text{Tr}[\Sigma(\theta)A]$

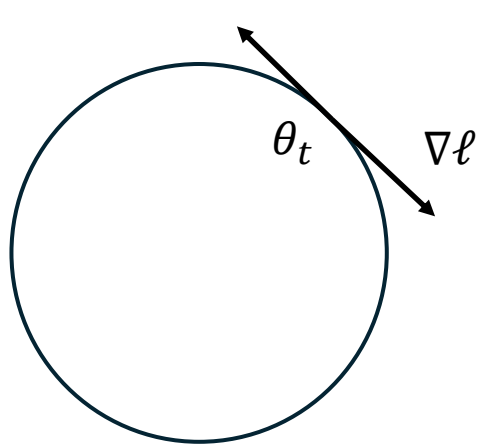
Theorem 4.3 (Fixed point theorem. Informal). For every A-exponential symmetry, and every θ , there exists a **unique** and **attractive** λ^* such that,

$$\frac{d}{dt} C_A(e^{\lambda^* A} \theta) = 0$$

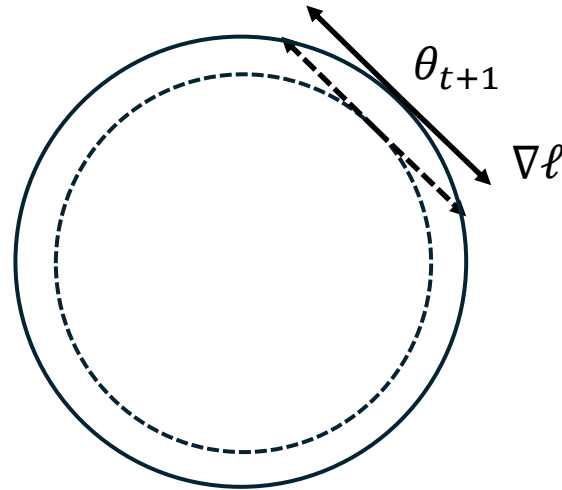
- Every exponential symmetry leads to a unique and attractive fixed point (in the degenerate direction) for training.

Example of Scale Invariant Losses

- Consider a 2d problem with scale invariance: $\ell(\theta) = \ell(\lambda\theta)$
- The gradient $\nabla\ell$ must be tangent to conservation laws



Step t



Step $t + 1$

Scale invariance \rightarrow
A systematic flow towards infinity

Applications

- A deep linear network:

$$\ell = |W_D W_{D-1} \dots W_2 W_1 x - y|^2$$

- Global minimum: $W_D W_{D-1} \dots W_2 W_1 = V^*$
- What is the fixed point of SGD?
 1. **Orthogonality**: W_{D-1}, \dots, W_2 are all (scalar multiples of) orthogonal matrices
 2. **Alignment**: $W_{D-1} \dots W_2$ aligns with the left singular matrix $\sqrt{\Sigma_\epsilon} W_D$ and the right singular matrix of $W_1 \sqrt{\Sigma_x}$
 3. **Balance**: all the following matrices have the same norm:
 $\sqrt{\Sigma_\epsilon} W_D, W_{D-1}, \dots, W_2, W_1 \sqrt{\Sigma_x}$

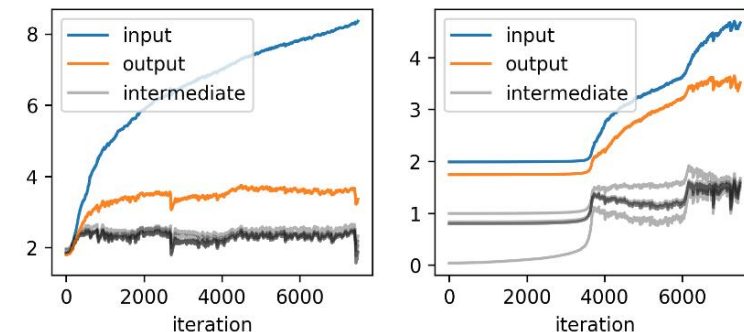


Figure 8: Norms of weights of multilayer deep linear network during training on MNIST without weight decay. We see that the intermediate layers converge to the same norm during training, whereas the input and output layers are different because they are determined by the input and output noise. This effect is robust against different initializations. This agrees with our analysis for deep linear nets (**Theorem 5.4**). **Left**: initializing all layers with the same norm. **Right**: initializing all layers at randomly different norms.

SGD can drive the sharpness both up and down (depending on Σ_x and Σ_ϵ and init.)

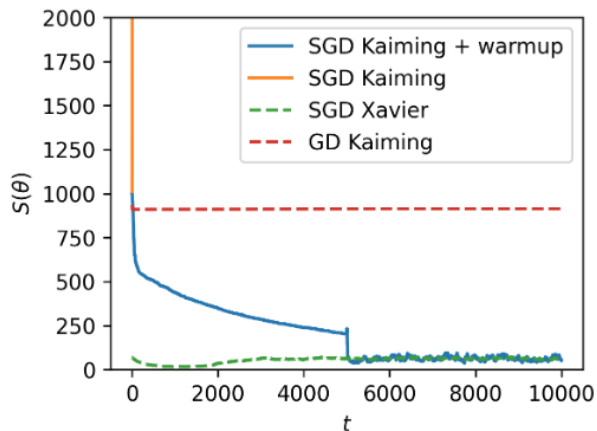


Figure 3: Dynamics of the stability condition S during the training of a rank-1 matrix factorization problem. The solid lines show the training of SGD with Kaiming init. When the learning rate ($\eta = 0.008$) is too large, SGD diverges (orange line). However, when one starts training at a small learning rate (0.001) and increases η to 0.008 after 5000 iterations, the training remains stable. This is because SGD training improves the stability condition during training, which is in agreement with the theory. In contrast, the stability condition of GD and that of SGD with a Xavier init increases only slightly. Also, note that both Xavier and Kaiming init. under SGD converges to the same stability condition because the equilibrium is unique.

Similar alignment in nonlinear nets

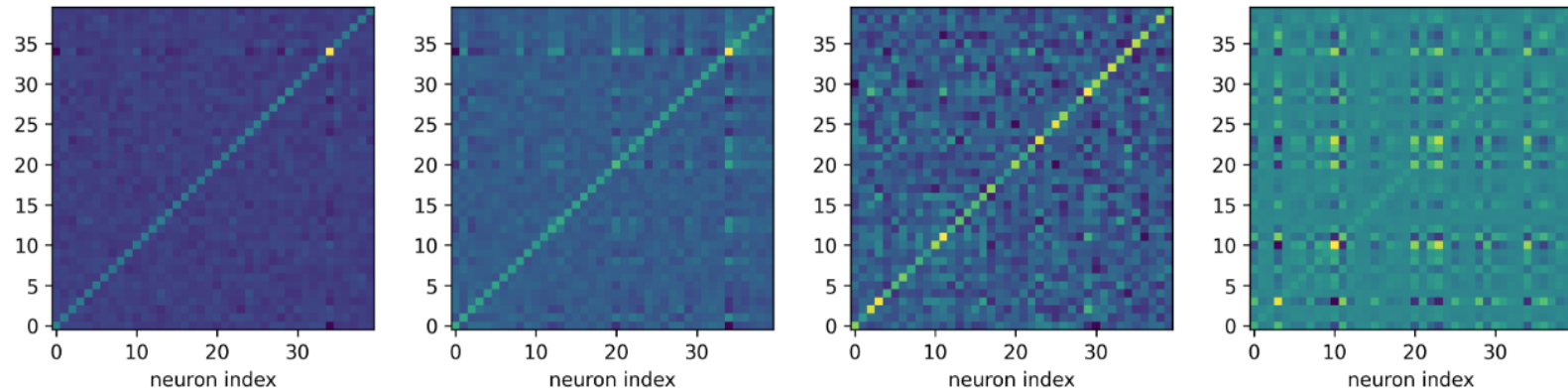


Figure 4: The latent representations of a two-layer tanh net trained under SGD (**left**) are similar across different layers, in agreement with the theory. However, the learned representations are dissimilar under GD (**right**). Here, we plot the matrices $W\bar{\Sigma}_xW$ (first and third plots) and $U\bar{\Sigma}_\epsilon U$ (second and fourth plots). Note that the quantity $W\bar{\Sigma}_xW$ is equal to the covariance of the preactivation representation of the first layer. This means that SGD and GD learn qualitatively different features after training. Also, see Appendix A.3 for other activations.

Message

- Symmetry implies different things for deterministic dynamics and stochastic dynamics
 - Deterministic: conservation law
 - Stochastic: unique equilibrium
- Under GD, the learned solutions are determined by initialization
- Under SGD, the learned solutions are determined by the gradient noise, and **independent of initialization**

[2] Loss Symmetry and Noise Equilibrium of Stochastic Gradient Descent. *NeurIPS 2024*

Table of contents

1. Physics of deep learning?
2. Discrete symmetry + Regularization = Constrained Parameters
3. Continuous Symmetry + Noise = Init. Independent Solutions
4. Noise + Regularization = Compact Representations

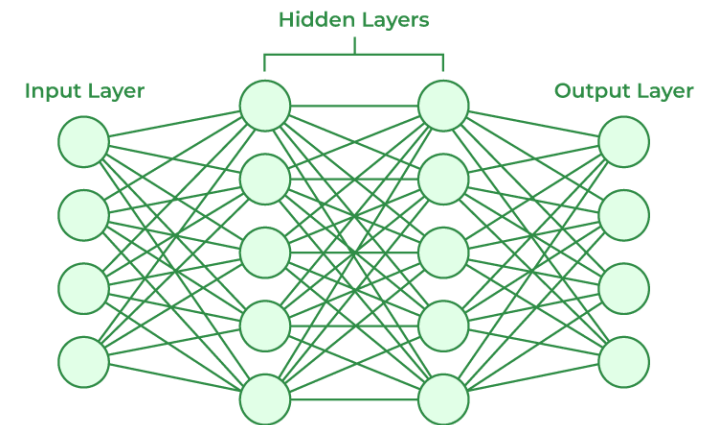
[3] Formation of Representations in Neural Networks.
[arxiv/2410.03006](https://arxiv.org/abs/2410.03006)

Latent Representation of Neural Networks

- Neural Networks process information layer by layer:
$$x \rightarrow h_1 \rightarrow h_2 \rightarrow \dots \rightarrow h_D \rightarrow y$$
- During training, the latent variables h becomes increasingly structured
- Of particular interest to deep learning and neuroscience is the second moment of the latent variables

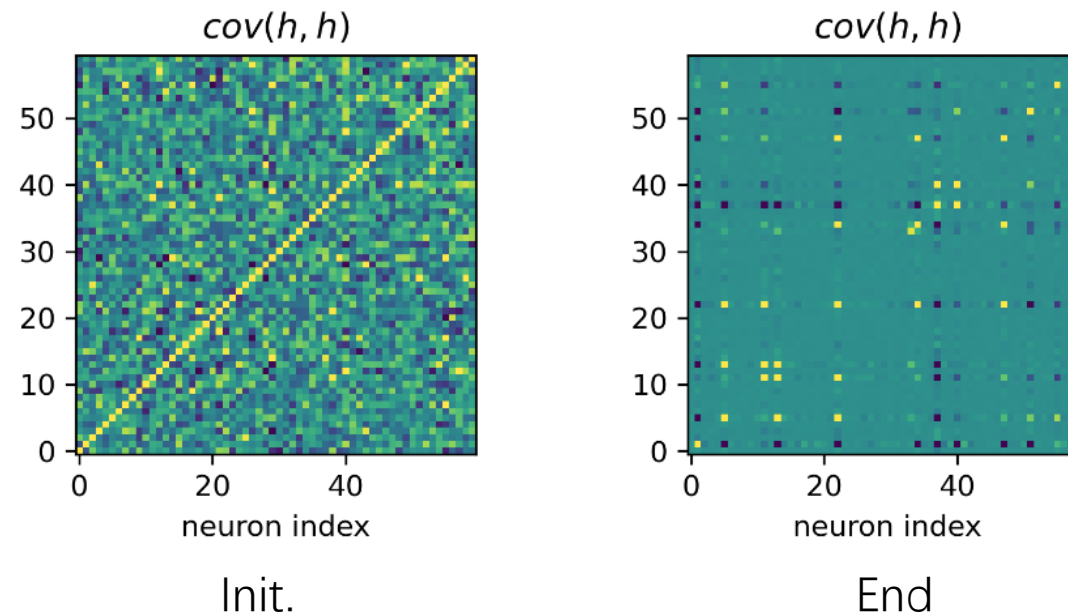
$$H := \mathbb{E}[hh^T]$$

- We will refer to H as the **representation**



Latent Representation

- After training, the representation becomes highly structured:



Latent representation

- Consider any layer connected by a linear weight:

$$h_b = Wh_a(x)$$

- The model is arbitrarily nonlinear:

$$f(x) = f(h_b(x))$$

- We are also concerned with the gradient of the neurons

$$g_b = \nabla_{h_b} \ell$$

$$g_a = \nabla_{h_a} \ell$$

- $\ell(\theta, x)$ is the per-sample loss

How does latent representations form?

- **Neural Collapse (NC)** is found to happen in the penultimate layer of overparametrized classifiers

- When NC happens,

$$\mathbb{E}[h_a h_a^T] \propto W^T W$$

Papayan et al.
PNAS 2020

- **Neural feature ansatz (NFA)** states that during the training of fully connected networks,

$$W^T W \propto \mathbb{E}[\nabla_{h_a} \ell (\nabla_{h_a} \ell)^T]$$

Radhakrishnan et al.
Science 2024

Canonical Representation Hypothesis

- Together, this implies that the neuron gradient g , weight W , and activations h are well aligned
- There exists six possible alignments between these quantities of the same layer:

representation-gradient alignment (RGA): $H_c \propto G_c$,

representation-weight alignment (RWA): $H_c \propto Z_c$,

gradient-weight alignment (GWA): $G_c \propto Z_c$,

where $c \in \{a, b\}$, $H_c = \mathbb{E}[h_c h_c^T]$, $G_c = \mathbb{E}[g_c g_c^T]$, $Z_a = W^T W$, $Z_b = W W^T$

- That all six relations are satisfied is referred to as the **CRH**

Canonical Representation Hypothesis

- There exists six possible alignments between these quantities of the same layer:

representation-gradient alignment (RGA): $H_c \propto G_c$,

representation-weight alignment (RWA): $H_c \propto Z_c$,

gradient-weight alignment (GWA): $G_c \propto Z_c$,

where $c \in \{a, b\}$, $H_c = \mathbb{E}[h_c h_c^T]$, $G_c = \mathbb{E}[g_c g_c^T]$, $Z_a = W^T W$, $Z_b = W W^T$

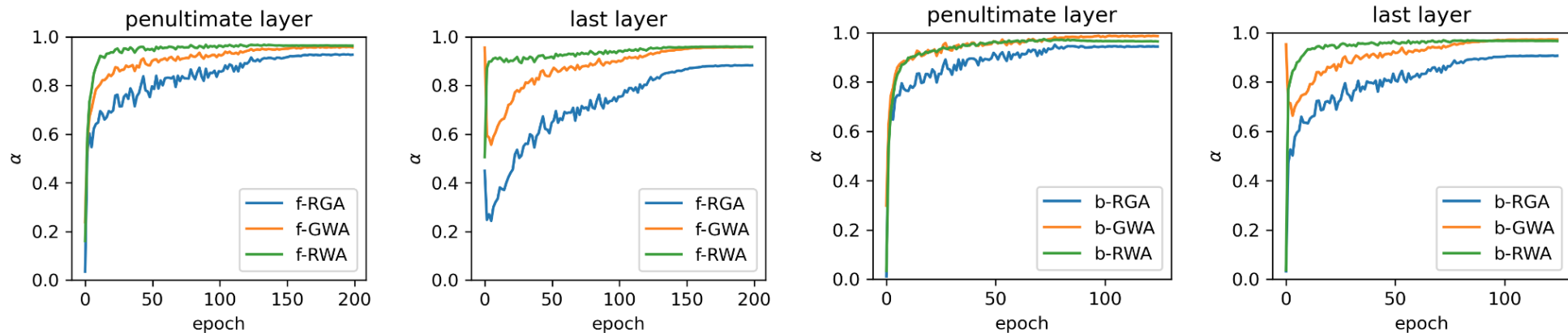
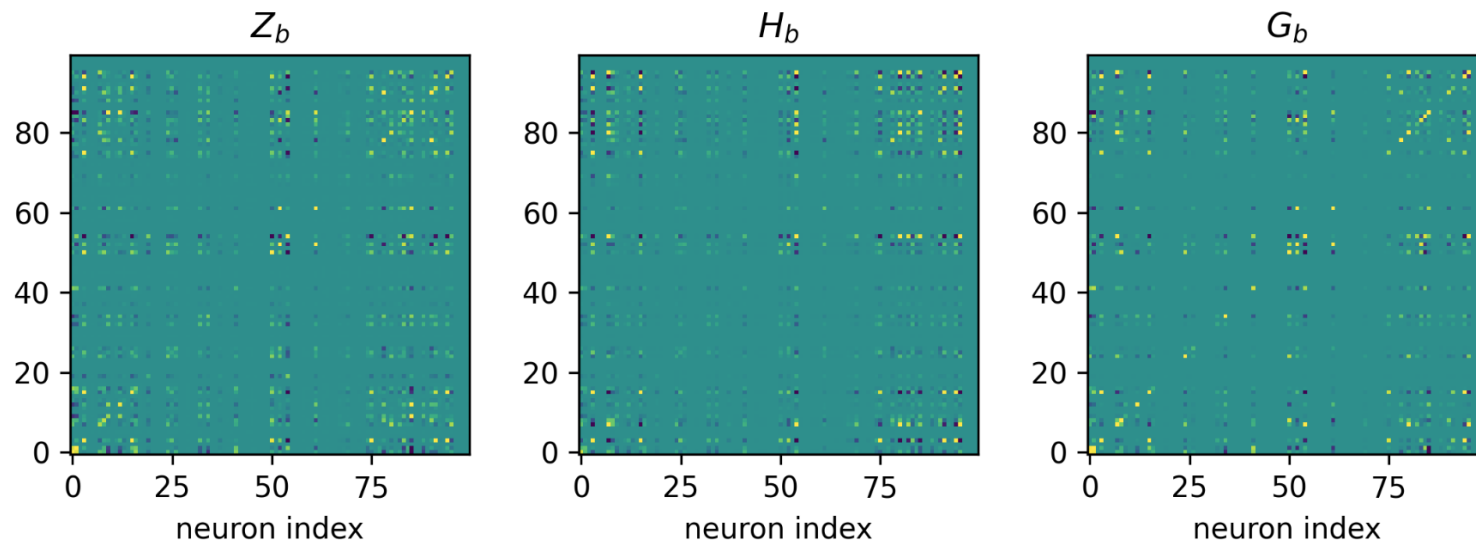


Figure 1: Six alignment relations in the penultimate layer and output layer of a ResNet18 trained on CIFAR-10 (**res1**). **Left:** forward CRH. **Right:** backward CRH. We see that all six relations hold significantly across two fully connected layers. Also, we show that the matrix $\text{cov}(g, h)$ is well aligned with WW^\top in the appendix Section [C.7](#), which is a strong piece of evidence supporting the key theoretical step that the cross terms will be aligned with the weights (and G , H).



Canonical Representation Hypothesis

- Meaning:

1. **Neuron variation** is aligned with its **importance**
2. Representation is fully **compact** and **robust to perturbations**
3. The information processing becomes invertible

How to prove it?

- Consider an overdamped particle moving in a harmonic potential, with random force $\sqrt{D}\xi(t)$:

$$\dot{x} = -\mu x + \sqrt{D}\xi(t)$$



How to prove it?

- Consider an overdamped particle moving in a harmonic potential, with random force $\sqrt{D}\xi(t)$:

$$\dot{x} = -\mu x + \sqrt{D}\xi(t)$$

- At stationarity, the rate of dissipation must balance with the fluctuation:

$$D = \mu \langle x^2 \rangle$$

- This is the *Einstein Relation (1905)*

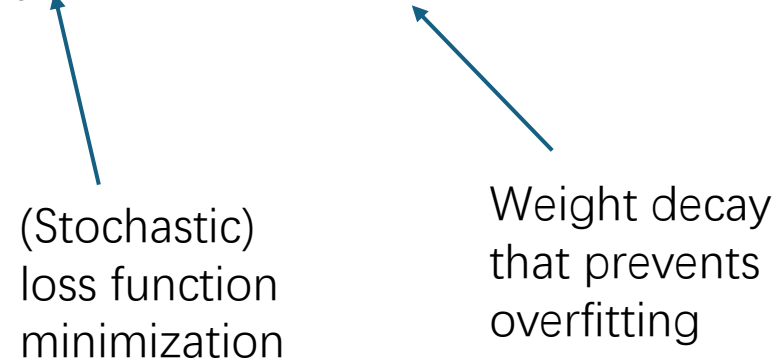


Fluctuation-Dissipation Theorem

- In deep learning, the time evolution is induced by the learning algorithm (SGD):

$$\Delta\theta = -\eta(\nabla_{\theta}\ell(\theta, x) - \gamma\theta)$$

- η is the learning rate



(Stochastic)
loss function
minimization

Weight decay
that prevents
overfitting

Fluctuation-Dissipation Theorem

- For our purpose, we are interested in the time evolution of the representation:

$$\Delta(h_b(x)h_b^\top(x)) = \eta(\|h_a\|^2 g_b h_b^\top + \|h_a\|^2 h_b g_b^\top - 2\gamma h_b h_b^\top) + \eta^2 \|h_a\|^4 g_b g_b^\top + O(\eta^2 \gamma + \|\Delta(h_a h_a^\top)\|),$$

- At stationarity,

$$0 = \underbrace{z_b \mathbb{E}[g_b h_b^\top] + z_b \mathbb{E}[h_b g_b^\top]}_{\text{learning}} - \underbrace{2\gamma \mathbb{E}[h_b h_b^\top]}_{\text{regularization}} + \underbrace{\eta z_b^2 \mathbb{E}[g_b g_b^\top]}_{\text{noise}}, \quad (6)$$

Drift, Energy minimization

Diffusion

Fluctuation–Dissipation Theorem

Theorem 1. Under Assumption 1, when $\mathbb{E}[\Delta(h_a h_a^\top)] = 0$, $\mathbb{E}[\Delta(g_b g_b^\top)] = 0$, $\mathbb{E}[\Delta(WW^\top)] = 0$, and $\mathbb{E}[\Delta(W^\top W)] = 0$, there exist real-valued constants $c_1, c_2, c_3, c_4 > 0$ such that

$$WW^\top + c_1 \mathbb{E}[g_b g_b^\top] = c_2 \mathbb{E}[h_b h_b^\top], \quad W^\top W + c_3 \mathbb{E}[h_a h_a^\top] = c_4 \mathbb{E}[g_a g_a^\top]. \quad (7)$$

Additionally, if at a local minimum,

$$WW^\top \propto \mathbb{E}[g_b g_b^\top] \propto \mathbb{E}[h_b h_b^\top], \quad W^\top W \propto \mathbb{E}[h_a h_a^\top] \propto \mathbb{E}[g_a g_a^\top]. \quad (8)$$

Polynomial Alignment Hypothesis

- The spectra of H , G , Z are power-laws of each other when the CRH is violated

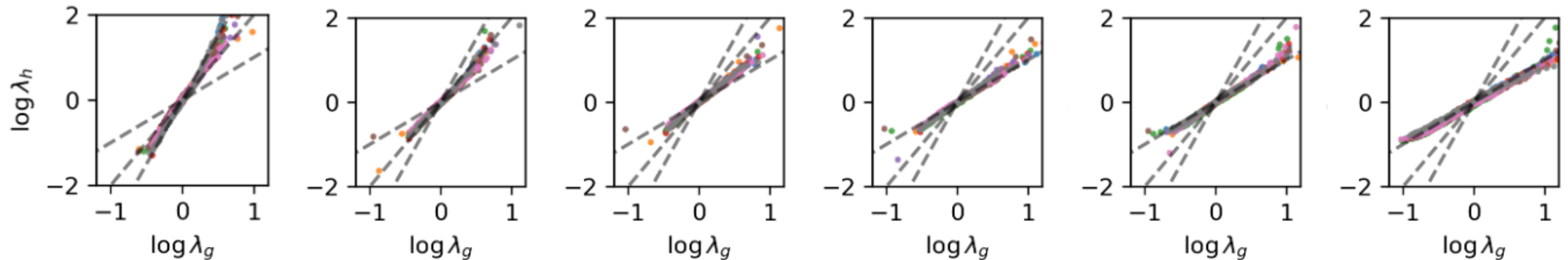


Figure 3: The power-law alignment between the eigenvalues λ_h and λ_g of H_b and G_b in a six-hidden layer transformer (**llm**). Left to Right: first to the penultimate layers. The grey dashed lines show the power-law relations $\lambda_h \propto \lambda_g^\alpha$ for $\alpha = 1, 2, 3$ respectively. We see that the first layer has an exponent of 3, the second has an exponent of 2, and all the layers after it are observed to have an exponent of 1. Different colors show different heads within the same layer. The range of the power exponents is in almost perfect agreement with the predicted range in Table [1](#). Referring to the table, this implies that these layers are in phases 5, 8, and 6, respectively. The setting is the same as the LLM experiment. Also, see Section [C.8](#) for fully connected nets.

Summary

- Fluctuation-Dissipation Theorem (FDT) is a main mechanism for the formation of representation in neural networks
- FDT leads to the emergence of the canonical representation, where weights, representation and gradients are all aligned
- Breaking of the CRH leads to reciprocal power-law relations

Takeaways

- Can we understand fundamental aspects of deep learning with physics?
 - **Yes**
- Are there universal laws (“alternative physics”) that exists uniquely in deep learning?
 - **Yes (and we need to invent new mathematical tools to find them!)**
- How can physics help understand deep learning?

How can physics help understand deep learning?

- Treat deep learning as an **empirical science**
 - “This Analysis consists in making Experiments and Observations, and in drawing general Conclusions from them by Induction” -- Newton
- Connect concepts from physics to concepts in deep learning
 - E.g. symmetry = constraint
- Use physics mechanisms to explain deep learning phenomena
 - E.g. Fluctuation Dissipation Theorem → Representation alignment
- Leverage these understanding to design novel algorithms
 - E.g. engineer artificial symmetries to compress neural networks

Collaborators



Isaac Chuang



Tomaso Poggio



Hongchao Li



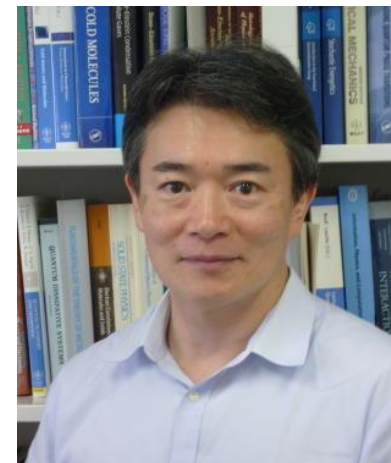
Mingze Wang



Tomer Galanti



Lei Wu



Masahito Ueda

Thanks

- [1] Symmetry Induces Structure and Constraint of Learning. *ICML 2024*
- [2] Loss Symmetry and Noise Equilibrium of Stochastic Gradient Descent. *NeurIPS 2024*
- [3] Formation of Representations in Neural Networks. [arxiv/2410.03006](https://arxiv.org/abs/2410.03006)

Violation of CRH

- CRH is more like to be violated than hold
- The following theorem characterizes what happens if the CRH is partially violated:

Theorem 2 (CRH Master Theorem). *Let A, B, C be a permutation of $\mathbb{E}[hh^\top]$, $\mathbb{E}[gg^\top]$, and Z , and let $\tilde{D} := PDP$ be a projected version of D for a projection matrix P . Then,*

1. *(Directional Redundancy) if any two forward (backward) alignments hold, all forward (backward) alignments hold;*
2. *(Reciprocal Polynomial Alignments) if one of any forward alignments and one of any backward alignments hold, there exists scalars α_c, β_c , and δ_c satisfying $-1 \leq \alpha_c, \beta_c, \delta_c \leq 3$ such that*

$$\tilde{A}_c^{\alpha_c} \propto \tilde{B}_c^{\beta_c} \propto \tilde{C}_c^{\delta_c}, \quad (9)$$

(as detailed in Table [1](#)) where $c \in \{a, b\}$ denotes the backward and forward relations respectively, and the corresponding projection $P_c \in \{Z_c^0, \mathbb{E}[h_c h_c^\top]^0, \mathbb{E}[g_c g_c^\top]^0\}$, e.g. such that $\tilde{A} = P_c A P_c$.

3. *(Canonical Alignment I) If (any) one more relation holds in addition to part 2, then all six alignments hold in the Z^0 subspace; in addition, at a local minimum, all six alignments hold;*
4. *(Canonical Alignment II) If all six alignments hold, $\mathbb{E}[hh^\top] \propto \mathbb{E}[gg^\top] \propto Z \propto P$, where P is an orthogonal projection matrix.*

Polynomial Alignment Hypothesis

Phase	Back. Alignment	Forw. Alignment	Back. Power Law	Forw. Power Law	NC	NFA	CU	llm
CRH	$H_a \propto Z_a \propto G_a$	$H_b \propto Z_b \propto G_b$	-	-	✓	✓	✓	
back. CRH	$H_a \propto Z_a \propto G_a$	-	-	$\tilde{H}_b^0 \propto \tilde{Z}_b^0 \propto \tilde{G}_b$ ($H_b \propto Z_b^2$)		✓	✓	
forw. CRH	-	$H_b \propto Z_b \propto G_b$	$\tilde{H}_a \propto \tilde{Z}_a^0 \propto \tilde{G}_a$ ($Z_a^2 \propto G_a$)	-				✓(3-6)
1	$H_a \propto G_a$	$H_b \propto G_b$	$\tilde{H}_a^0 \propto \tilde{Z}_a \propto \tilde{G}_a^0$	$\tilde{H}_b^0 \propto \tilde{Z}_b \propto \tilde{G}_b^0$				✓(3-6)
2	$H_a \propto Z_a$	$H_b \propto Z_b$	$\tilde{H}_a \propto \tilde{Z}_a \propto \tilde{G}_a^0$	$\tilde{H}_b \propto \tilde{Z}_b \propto \tilde{G}_b^0$			✓	
3	$G_a \propto Z_a$	$G_b \propto Z_b$	$\tilde{H}_a^0 \propto \tilde{Z}_a \propto \tilde{G}_a$	$\tilde{H}_b^0 \propto \tilde{Z}_b \propto \tilde{G}_b$		✓		
4	$H_a \propto G_a$	$H_b \propto Z_b$	$\tilde{H}_a \propto \tilde{Z}_a^0 \propto \tilde{G}_a$	$\tilde{H}_b \propto \tilde{Z}_b \propto \tilde{G}_b^{-1}$				
5	$H_a \propto Z_a$	$H_b \propto G_b$	$\tilde{H}_a^3 \propto \tilde{Z}_a^3 \propto \tilde{G}_a$	$\tilde{H}_b \propto \tilde{Z}_b^2 \propto \tilde{G}_b$			✓	✓(3-6)
6	$H_a \propto G_a$	$G_b \propto Z_b$	$\tilde{H}_a \propto \tilde{Z}_a^2 \propto \tilde{G}_a$	$\tilde{H}_b \propto \tilde{Z}_b^3 \propto \tilde{G}_b^3$				✓(1)
7	$G_a \propto Z_a$	$H_b \propto G_b$	$\tilde{H}_a^{-1} \propto \tilde{Z}_a \propto \tilde{G}_a$	$\tilde{H}_b \propto \tilde{Z}_b^0 \propto \tilde{G}_b$		✓		
8	$H_a \propto Z_a$	$G_b \propto Z_b$	$\tilde{H}_a^2 \propto \tilde{Z}_a^2 \propto \tilde{G}_a$	$\tilde{H}_b \propto \tilde{Z}_b^2 \propto \tilde{G}_b^2$			✓	✓(2)
9	$G_a \propto Z_a$	$H_b \propto Z_b$	$\tilde{H}_a \propto \tilde{Z}_a^0 \propto \tilde{G}_a^0$	$\tilde{H}_b^0 \propto \tilde{Z}_b^0 \propto \tilde{G}_b$		✓		

Table 1: The reciprocal polynomial relations of the CRH Master Theorem. When one forward relation and one backward relation hold simultaneously, all six matrices are polynomially aligned in a subspace (Theorem 2). Each scaling relationship can be regarded as a possible phase for the layer during actual training. The right panel shows how existing observations about neural networks fit into the phase diagram. A ✓ denotes that this phenomenon is compatible with the specified phase. NC refers to the neural collapse. NFA refers to the neural feature ansatz. CU (correlated update) refers to the (idealization of the) common observation that h_a is correlated with W a few steps after training (Everett et al., 2024). The llm column shows the compatibility of the scaling relation for transformer observed in Figure 3.

Polynomial Alignment Hypothesis

- The spectra of H , G , Z are power-laws of each other when the CRH is violated

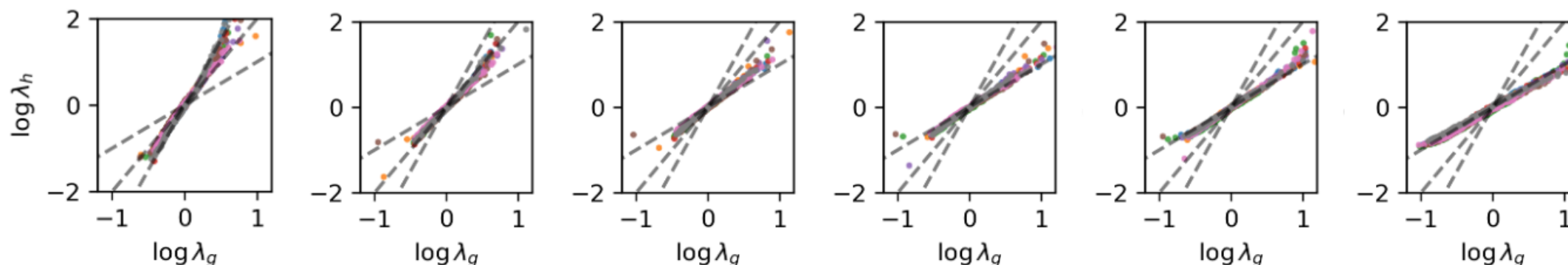


Figure 3: The power-law alignment between the eigenvalues λ_h and λ_g of H_b and G_b in a six-hidden layer transformer (**llm**). Left to Right: first to the penultimate layers. The grey dashed lines show the power-law relations $\lambda_h \propto \lambda_g^\alpha$ for $\alpha = 1, 2, 3$ respectively. We see that the first layer has an exponent of 3, the second has an exponent of 2, and all the layers after it are observed to have an exponent of 1. Different colors show different heads within the same layer. The range of the power exponents is in almost perfect agreement with the predicted range in Table [1](#). Referring to the table, this implies that these layers are in phases 5, 8, and 6, respectively. The setting is the same as the LLM experiment. Also, see Section [C.8](#) for fully connected nets.