ICML
International Conference
On Machine Learning

# On the Emergence of Cross-Task Linearity in the Pretraining-Finetuning Paradigm

**Zhanpeng Zhou**[1*], Zijun Chen[1,2*], Yilan Chen[3], Bo Zhang[2§], Junchi Yan[1,2§*]

[1]Shanghai Jiao Tong University, [2]Shanghai AI Lab, [3]University of California San Diego

SHANGHAI JIAO TONG UNIVERSITY

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

UC San Diego

*Equal contribution, §Corresponding author

# Background: LMC

**Linear Mode Connectivity (LMC)**

Given dataset $D$ and two modes $\boldsymbol{\theta}_A, \boldsymbol{\theta}_B$ that $\mathrm{Err}_D(\boldsymbol{\theta}_A) = \mathrm{Err}_D(\boldsymbol{\theta}_B)^*$, two mode $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ satisfy the *linear mode connectivity* if

$$\forall \alpha \in [0,1], \mathrm{Err}_D(\alpha\boldsymbol{\theta}_A + (1-\alpha)\boldsymbol{\theta}_B) \approx \mathrm{Err}_D(\boldsymbol{\theta}_A)$$

*$\mathrm{Err}_D(\boldsymbol{\theta})$ denotes the classification error of the network $f(\boldsymbol{\theta};\cdot)$ on the dataset $D$.

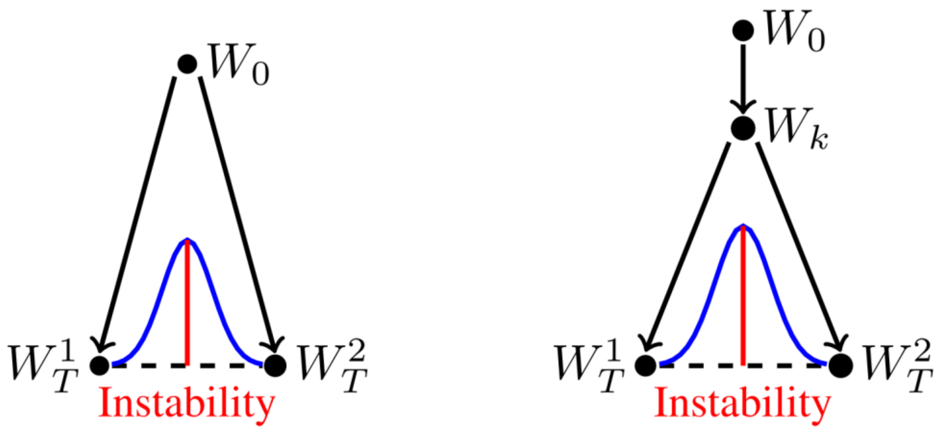



Fig. 1: Illustration of spawning method and LMC [1].

Frankle et al. [1] observed LMC for networks that are jointly trained for a short time before independent training (**spawning method**).

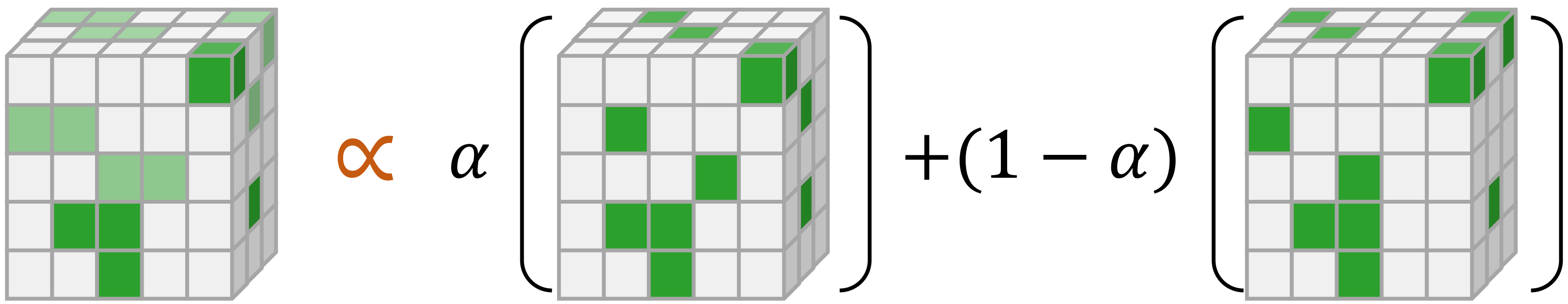
[1] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis.

# Background: LLFC

**Layerwise Linear Feature Connectivity (LLFC)**

Given dataset $D$ and two modes $\boldsymbol{\theta}_A$, $\boldsymbol{\theta}_B$ of an $L$-layer neural network $f$, the modes $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ are *layerwise linearly feature connected* if:

$$\forall \ell \in [L], \forall \alpha \in [0,1], \exists c > 0, s.t., c f^{(\ell)}(\alpha \boldsymbol{\theta}_A + (1-\alpha)\boldsymbol{\theta}_B) = \alpha f^{(\ell)}(\boldsymbol{\theta}_A) + (1-\alpha) f^{(\ell)}(\boldsymbol{\theta}_B).$$




$f^{(\ell)}(\alpha \boldsymbol{\theta}_A + (1-\alpha)\boldsymbol{\theta}_B; \boldsymbol{X})$ $\quad$ $f^{(\ell)}(\boldsymbol{\theta}_A; \boldsymbol{X})$ $\quad$ $f^{(\ell)}(\boldsymbol{\theta}_B; \boldsymbol{X})$

# Background: LLFC connects to LMC

**LLFC always co-occurs with LMC in practice**

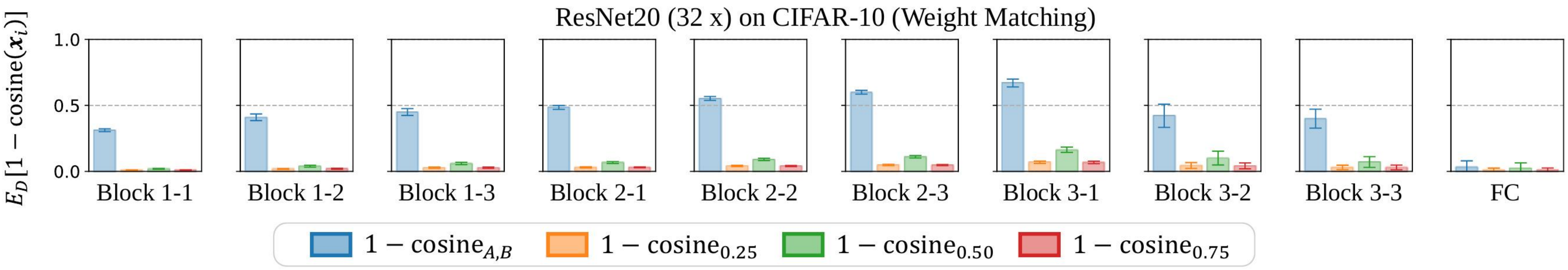



Fig. 2: Comparison of $E_D[1 - \mathrm{cosine}_\alpha(\boldsymbol{x}_i)]^*$ and $E_D[1 - \mathrm{cosine}_{A,B}(\boldsymbol{x}_i)]^*$, $\alpha \in \{.25, .5, .75\}$.

**Lemma (LLFC implies LMC)**

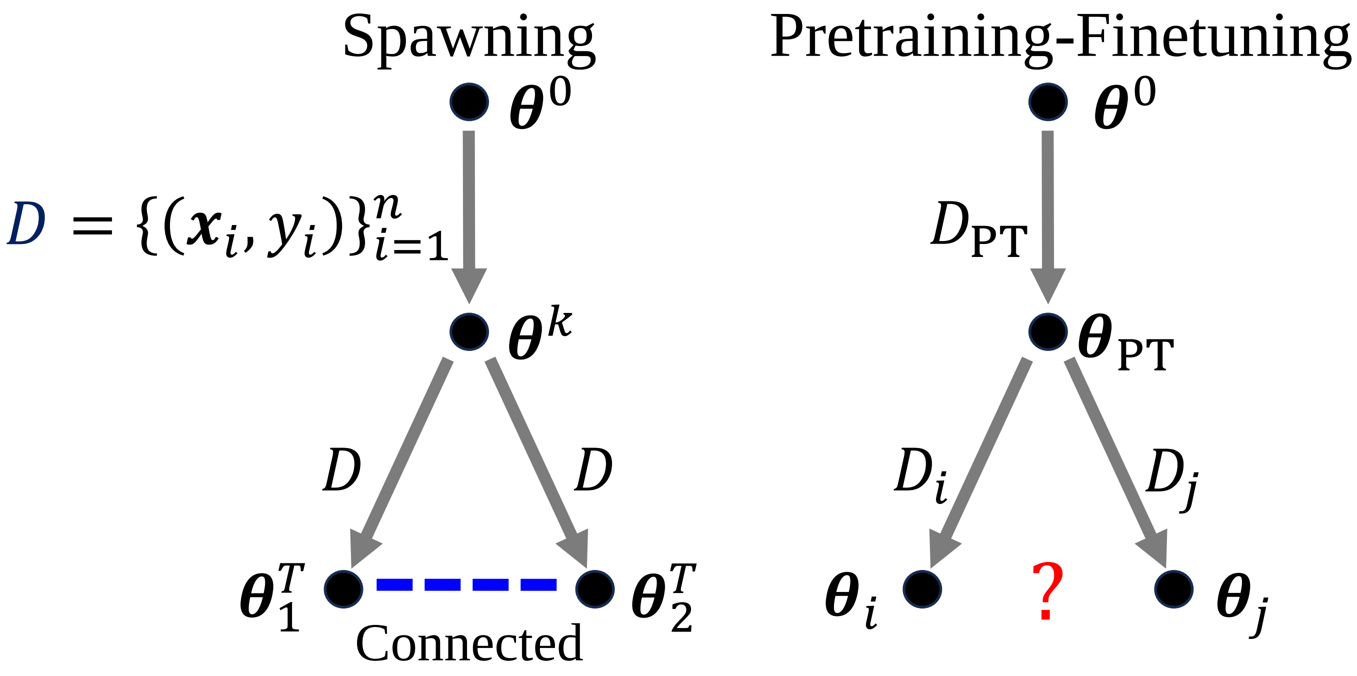
Two modes $\boldsymbol{\theta}_A$, $\boldsymbol{\theta}_B$ satisfy LLFC over dataset $D$ and $\max\{\mathrm{Err}_D(\boldsymbol{\theta}_A), \mathrm{Err}_D(\boldsymbol{\theta}_B)\} \leq \epsilon$, then

$$\forall \alpha \in [0,1], \mathrm{Err}_D(\alpha\boldsymbol{\theta}_A + (1-\alpha)\boldsymbol{\theta}_B) \leq 2\epsilon.$$

$^*\mathrm{cosine}_\alpha(\boldsymbol{x}_i) = \cos\langle f^{(\ell)}(\alpha\theta_A + (1-\alpha)\theta_B; \boldsymbol{x}_i), \alpha f^{(\ell)}(\theta_A; \boldsymbol{x}_i) + (1-\alpha) f^{(\ell)}(\theta_B; \boldsymbol{x}_i)\rangle$ and $\mathrm{cosine}_{A,B}(\boldsymbol{x}_i) = \cos\langle f^{(\ell)}(\theta_A; \boldsymbol{x}_i), f^{(\ell)}(\theta_B; \boldsymbol{x}_i)\rangle$

# Pretraining-Finetuning Paradigm

Intuition: Finetuning shares similar training regime with the spawning method.




***Are finetuned models linearly connected in loss landscape or feature space?***

*Fine-tuning can be done on the parameters of original neural network, or on "adaptors" consist of far fewer parameters than the original model. We focus on the former case.

# Cross-Task Linearity

**LMC fails, LLFC holds.**

Indeed, a stronger version of LLFC is observed, called *Cross-Task Linearity (CTL)*. Given a pair of finetuned models $(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \in \Theta^2$ and downstream tasks $D_i$ and $D_j$ respectively, we say them satisfy CTL on $D_i \cup D_j$ if

$$\forall \ell \in [L], \forall \alpha \in [0,1], s.t., f^{(\ell)}\left(\alpha \boldsymbol{\theta}_i + (1-\alpha)\boldsymbol{\theta}_j\right) \approx \alpha f^{(\ell)}(\boldsymbol{\theta}_i) + (1-\alpha) f^{(\ell)}\left(\boldsymbol{\theta}_j\right).$$

**Conjecture (Transitivity of CTL.)**

Given models $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j, \boldsymbol{\theta}_k$. We have $(\boldsymbol{\theta}_i, \boldsymbol{\theta}_k)$ satisfy CTL if $(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ and $(\boldsymbol{\theta}_j, \boldsymbol{\theta}_k)$ satisfy CTL.

We can further apply CTL to explain *Model Soup [6]* and *Task Arithmetic [6]*.

[5] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time.
[6] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, Ali Farhadi. Editing Models with Task Arithmetic.

# Insights into Model Averaging

**Model Averaging (Uniform Model Soup)**

Considering a set of models $\Theta = \{\boldsymbol{\theta}_i\}_k$ that started from $\boldsymbol{\theta}_{PT}$ and finetuned on the same task $D_{FT}$ but with different hyperparameter configuration, model averaging is defined as

$$f\left(\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{\theta}_i\right).$$

**Connect model averaging and model ensemble**

A finer-grained characterization of the linear correlation between model averaging and logits ensemble is observed.

$$f^{(\ell)}\left(\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{\theta}_i\right) = \frac{1}{k}\sum_{i=1}^{k} f^{(\ell)}(\boldsymbol{\theta}_i)\,, \forall \ell \in [L].$$

# Insights into Model Averaging

**Theorem (CTL generalizes to multiple models.)**
Given dataset $D$ and a set of modes $\Theta$ where each pair of models $(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \in \Theta^2$ satisfies CTL on $D$, assuming transitivity of CTL, then for any $\{\boldsymbol{\theta}_i\}_{i=1}^k \in \Theta$ and $\{\alpha_i\}_{i=1}^k \in [0,1]$, subject to the constraint that $\sum_{i=1}^k \alpha_i = 1$,

$$f^{(\ell)}\left(\sum_{i=1}^{n} \alpha_i \boldsymbol{\theta}_i\right) = \sum_{i=1}^{n} \alpha_i f^{(\ell)}(\boldsymbol{\theta}_i), \forall \ell \in [L].$$

The connection between model averaging and ensemble can be viewed as a generalization of CTL to the case of multiple models in the pretraining-finetuning paradigm.

# Insights into Task Arithmetic

**Task Arithmetic**

Considering a set of modes $\Theta = \{\boldsymbol{\theta}_i\}_k$ that started from $\boldsymbol{\theta}_{PT}$ but finetuned on different tasks $\{D_i\}_k$, task vector $\{\tau_i\}_k$ is defined as $\tau_i = \boldsymbol{\theta}_i - \boldsymbol{\theta}_{PT}$. Arithmetic operations can be applied to task vectors to construct $\tau_{new}$ and $\tau_{new}$ can be applied to $\boldsymbol{\theta}_{PT}$, i.e.,

$$f(\boldsymbol{\theta}_{PT} + \lambda\tau_{new}).$$

**CTL explains learning via addition.**

$f\big(\boldsymbol{\theta}_{PT} + \lambda(\tau_i + \tau_j)\big)$ demonstrate abilities on both $D_i$ and $D_j$. As CTL holds (verified empirically), $\forall \ell \in [L]$,

$$f^{(\ell)}\big(\boldsymbol{\theta}_{PT} + \lambda(\tau_i + \tau_j)\big) \approx \frac{1}{2} f^{(\ell)}(\boldsymbol{\theta}_{PT} + 2\lambda\tau_i) + \frac{1}{2} f^{(\ell)}\big(\boldsymbol{\theta}_{PT} + 2\lambda\tau_j\big).$$

*Addition over parameter space can be transformed to feature space.*

# Insights into Task Arithmetic

**CTL explains forgetting via negation.**
$f(\boldsymbol{\theta}_{PT} - \lambda\tau_i)$ loses ability on $D_i$ while retains performance elsewhere. As CTL holds (verified empirically),

$$f^{(\ell)}(\boldsymbol{\theta}_{PT}) \approx \frac{1}{2} f^{(\ell)}(\boldsymbol{\theta}_{PT} - \lambda\tau_i) + \frac{1}{2} f^{(\ell)}\big(\boldsymbol{\theta}_{PT} + \lambda\tau_j\big).$$

We rewrite it as

$$f^{(\ell)}(\boldsymbol{\theta}_{PT} - \lambda\tau_i) \approx f^{(\ell)}(\boldsymbol{\theta}_{PT}) - \Delta^{(\ell)}(\lambda\tau_i),$$

where $\Delta^{(\ell)}(\lambda\tau_i) = f^{(\ell)}\big(\boldsymbol{\theta}_{PT} + \lambda\tau_j\big) - f^{(\ell)}(\boldsymbol{\theta}_{PT})$. Intuitively, $\Delta^{(\ell)}(\lambda\tau_i)$ encode the information specific to task $D_i$.

*Negation over parameter space can be transformed to feature space.*

# Unveiling the Root Cause of CTL

**Factors Contributing to CTL (Highlight the role of pretraining).**

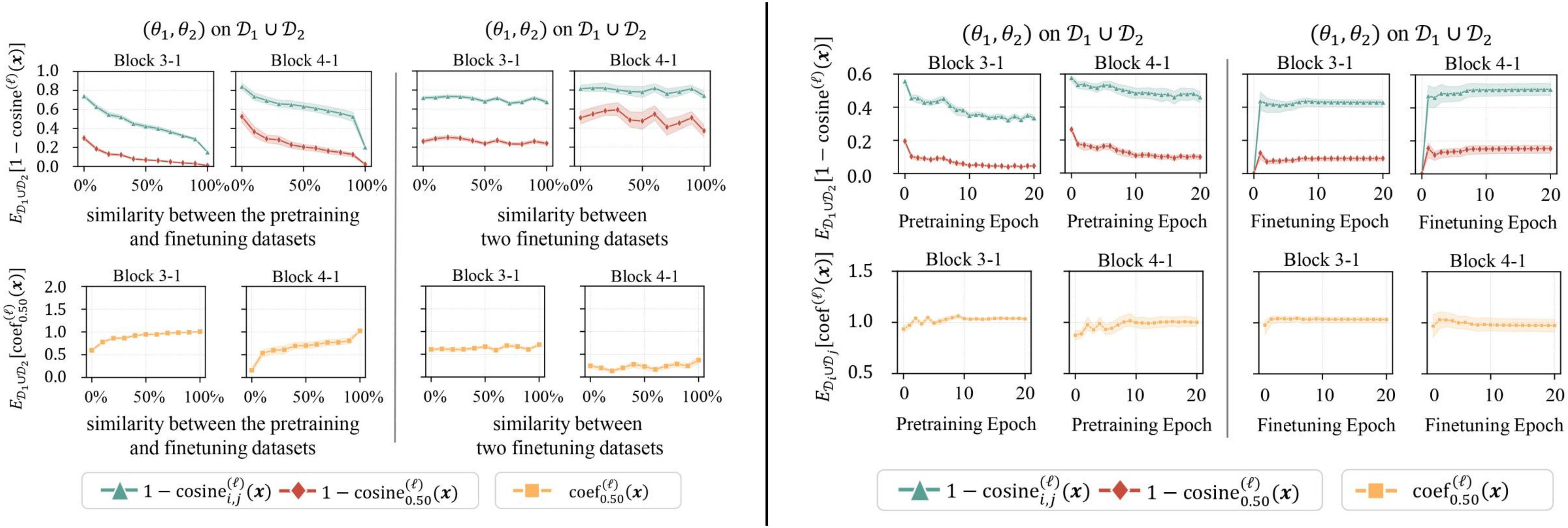



Fig. 3: The impact of the **task similarity (left)** /**number of pretraining and finetuning epochs (right)** on the emergence of CTL.

# Unveiling the Root Cause of CTL

**Theorem (The Emergence of CTL.)**
Suppose $\mathrm{f}(\boldsymbol{\theta}): R^p \mapsto R$ is third-differentiable function in an open convex set $\Theta$ and its Hessian norm at $\boldsymbol{\theta}_0$ is bounded by $\lambda_{min} \leq |\nabla^2 f(\boldsymbol{\theta}_0)| \leq \lambda_{max}$, then

$$|f\left(\alpha\boldsymbol{\theta}_i + (1-\alpha)\boldsymbol{\theta}_j\right) - \alpha f(\boldsymbol{\theta}_i) - (1-\alpha) f\left(\boldsymbol{\theta}_j\right)| \leq \frac{\alpha(1-\alpha)\lambda_{max}}{2} ||\boldsymbol{\theta}_i - \boldsymbol{\theta}_j||^2 + \epsilon,$$

Where $\epsilon = O(\max(||\alpha\boldsymbol{\theta}_i + (1-\alpha)\boldsymbol{\theta}_j - \boldsymbol{\theta}_0||^3, \alpha||\boldsymbol{\theta}_i - \boldsymbol{\theta}_0||^3, (1-\alpha)||\boldsymbol{\theta}_j - \boldsymbol{\theta}_0||^3))$ is the higher order term.

*Remarks:*

- *The emergence of CTL is related to the flatness of the function landscape and distance between two finetuned models.*
- *Instead of linearizing models, we provide a more realistic setting.*

# Thank you!

Q&A