Phase Diagram of Early Training Dynamics in Deep Networks:
Effect of the Learning Rate, Depth, and Width

Dayal Kalra and Maissam Barkeshli
NeurIPS 2023

Department of Physics,
University of Maryland, College Park (UMD)

Dec 12, 2023

Premise

Premise:

- We study the early training dynamics of DNNs trained using SGD with learning rate $\eta = {}^{c}/{\lambda_0^H}$.
  Here, c is a constant and λ_t^H is the top eigenvalue of the Hessian H (sharpness) at step t.
- By monitoring loss and sharpness, we study the effect of learning rate constant c, depth d, and width w on the early training dynamics

Architectures: Results validated FCNs, CNNs, and ResNets with ReLU activation and initial weight variance $\sigma_w^2 = {}^{2}/{\text{fan}_{\text{in}}}$.

Loss functions and datasets: MSE and cross-entropy; CIFAR-10, MNIST, and Fashion-MNIST.

The four regimes of neural network training

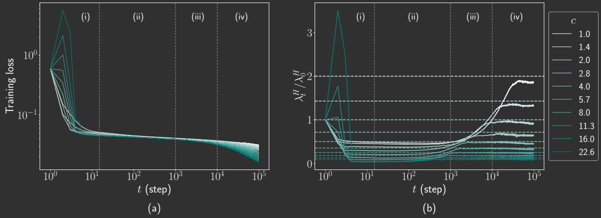


Figure: Training trajectories of CNNs trained on CIFAR-10 using MSE.

Typical training trajectories of deep networks show four training regimes:

- i **Early time transient:** loss and sharpness may drastically change depending on the learning rate and training eventually settles down.
- ii **Intermediate saturation:** sharpness plateaus before gradually increasing. (for the analysis of the intermediate saturation regime, refer to the paper)
- iii **Progressive sharpening:** sharpness increases until it reaches $\lambda^H \approx 2/\eta$ (Jastrzebski et. al 2020)
- iv **Late-time dynamics (EoS):** For MSE loss, sharpness oscillates around $2/\eta$. For cross-entropy loss, sharpness decreases after reaching $2/\eta$ (Cohen et. al 2021).

Early training dynamics of wide networks

Classical intuition from convex optimization requires $\eta\lambda_0^H = c < 2$.

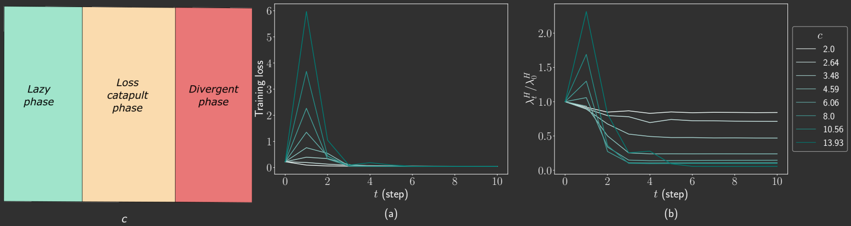


Figure: Early training dynamics of wide CNNs trained on CIFAR-10 using MSE.

Wide networks trained with MSE loss have three phases of early training wrt learning rate (Lewkowycz et al. 2020):

- **Lazy phase** $(c < 2)$: Loss monotonically decreases, sharpness remains constant
- **Catapult phase** $(2 < c < c_{max})$: Loss spikes initially, training converges with an abrupt decrease in sharpness
- **Divergent phase** $(c_{max} < c)$: Training diverges

Early training dynamics of deep networks

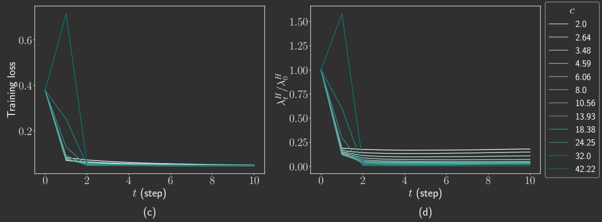


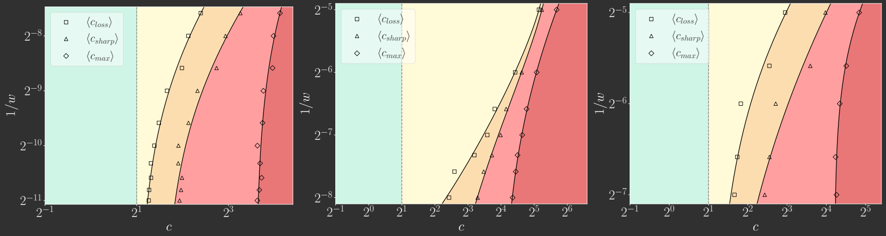
Figure: Early training dynamics of deep CNNs trained on CIFAR-10 using MSE.

Observation: For deep networks, training loss and sharpness may catapult only near the largest trainable learning rate.

To quantify the early training dynamics, define the following critical constants:

- (c_{loss}): Smallest learning rate constant resulting in early loss increase
- (c_{sharp}): Smallest learning rate constant resulting in early sharpness increase
- (c_{max}): Largest trainable learning rate constant during early training

Phase diagram of early training with width



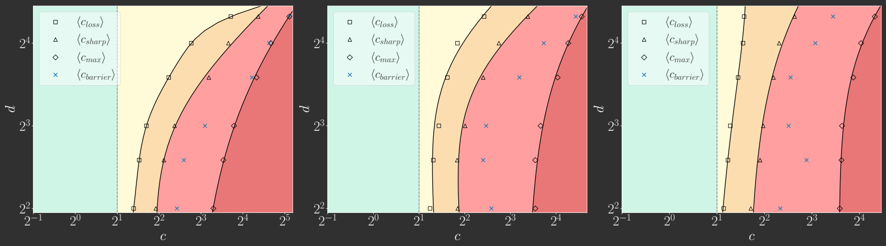
(a) FCNs on MNIST (b) CNNs on Fashion-MNIST (c) ResNets on CIFAR-10

Figure: Phase diagrams of early training of three different types of neural networks. Each data point $\langle c \rangle$ is an average over ten random initializations.

Observations:

- Critical constants $\langle c_{loss} \rangle, \langle c_{sharp} \rangle$, and $\langle c_{max} \rangle$ increase with $1/w$.
- In particular, $\langle c_{loss} \rangle$ deviates from $c = 2$ towards $\langle c_{max} \rangle$ on increasing in $^1/_w$.

Phase diagram of early training with depth



(a) FCNs $w = 512$ (b) FCNs $w = 1024$ (c) FCNs $w = 2048$

Figure: Phase diagrams of early training with depth for FCNs trained on Fashion-MNIST. Each data point $\langle c \rangle$ is an average over ten random initializations.

Observation: Similar phase diagrams emerge on replacing d with $^1/_w$.

The phase diagram of early training

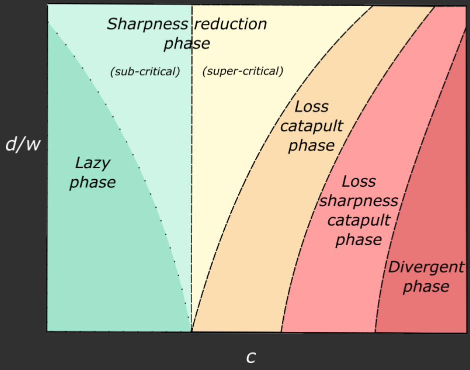
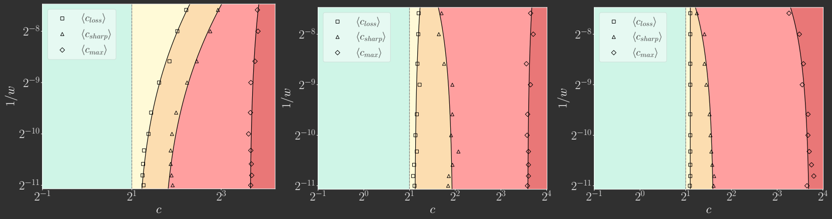


Figure: Sketch of the phase diagram of early training

Effect of network output at initialization

We examine the effect of network output by setting it to zero at initialization, $f(x;\theta_0) = 0$ by

1. centering the network $f_c(x;\theta) = f(x;\theta) - f(x;\theta_0)$
2. setting the last layer weights to zero at initialization



(a) vanilla (b) centered (c) last layer set to zero

Figure: Remarkably, both (1) and (2) remove the opening up of the sharpness reduction phase with $1/w$ and d.

Insights from a simple model

Definition

(uv model): Consider a two-layer linear network

$$f(x) = \frac{1}{\sqrt{w}} v^T u\, x, \qquad\qquad x, f \in \mathbb{R}$$

trained on a single training example $(x, y) = (1, 0)$ using MSE loss.

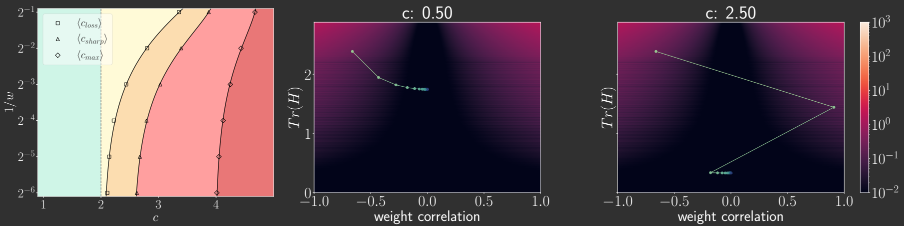


Figure: (left) uv model trained on a single example $(x, y) = (1, 0)$ exhibits a similar phase diagram. (right) training trajectories of uv model with $w = 2$ in a two-dimensional space defined by $Tr(H)$ and weight correlation $\cos(u, v)$

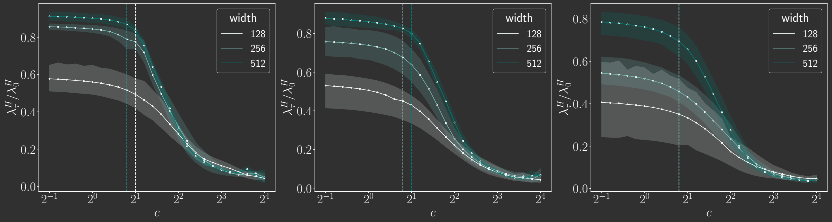
Thank You

Thank You!

https://openreview.net/forum?id=Al9yglQGKj

https://github.com/dayal-kalra/early-training

Abrupt reduction in sharpness with learning rate



(a) depth $= 5$ (b) depth $= 7$ (c) depth $= 10$

Definition

($\langle c_{crit} \rangle$) Given the averaged normalized sharpness $\langle \frac{\lambda^H_\tau}{\lambda^H_0} \rangle$ estimated using sharpness measured at τ, we define c_{crit} as

$$\langle c_{crit} \rangle = \arg\min_c \frac{\partial^2}{\partial c^2} \left\langle \frac{\lambda^H_\tau}{\lambda^H_0} \right\rangle \tag{1}$$

Observation: $\langle c_{crit} \rangle \approx 2$, irrespective of depth and width.

Insights from the training trajectories

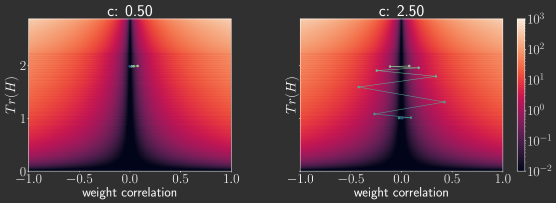


Figure: uv model with large width ($w = 512$).

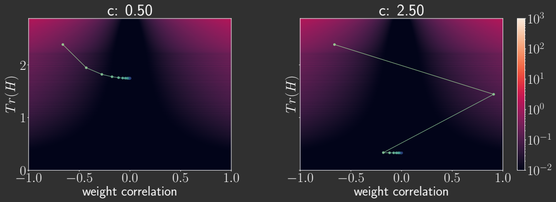


Figure: uv model with small width ($w = 2$).

Key References I