Infinite-width 2LP

High-dimensional 2LP

Deep ResNets

Large L, M, D limit

Conclusion

ResNets of All Shapes and Sizes

Louis-Pierre Chaintron (EPFL)

ProbAI Theory of Scaling Laws Workshop 2026
– University of Warwick, June 23rd

Most of this material is borrowed from Lénaïc Chizat.

Scaling up Neural Networks

Infinite-width 2LP

High-dimensional 2LP

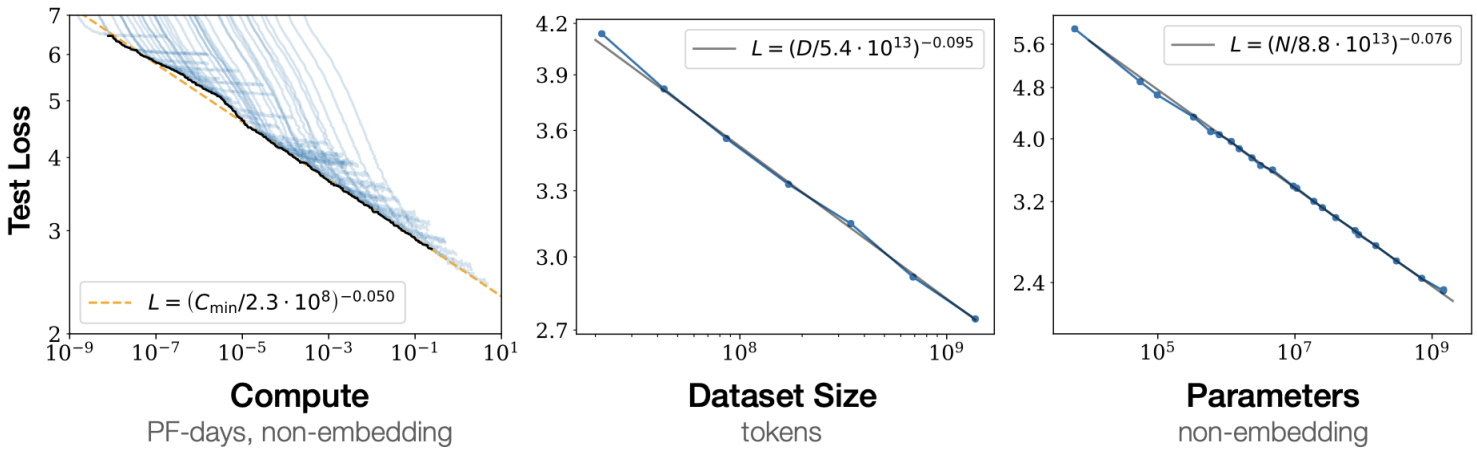
Deep ResNets

Large L, M, D limit

Conclusion

Observation:

- Scaling up compute budget of training NNs improves performance.
- Several ways : more data, longer training, bigger models.



Performance vs compute [Kaplan et al'20]
N parameters, dataset size D, compute C

$\hookrightarrow$ Classify large-scale limits wrt hyper-parameter (HP) scalings.

$\Rightarrow$ Improve choices of HP.

Residual Neural Networks

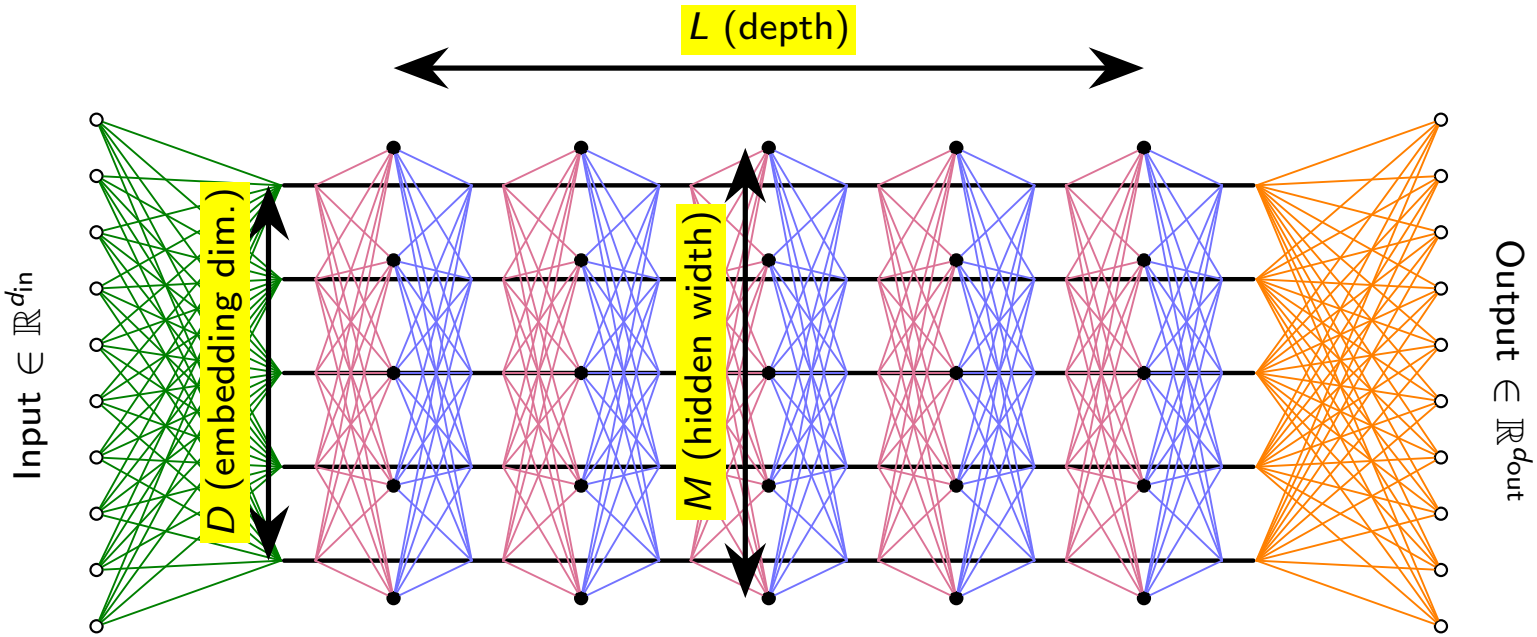
Infinite-width 2LP

High-dimensional 2LP

Deep ResNets

Large L, M, D limit

Conclusion



Application: HP transfer [Yang et al'22]

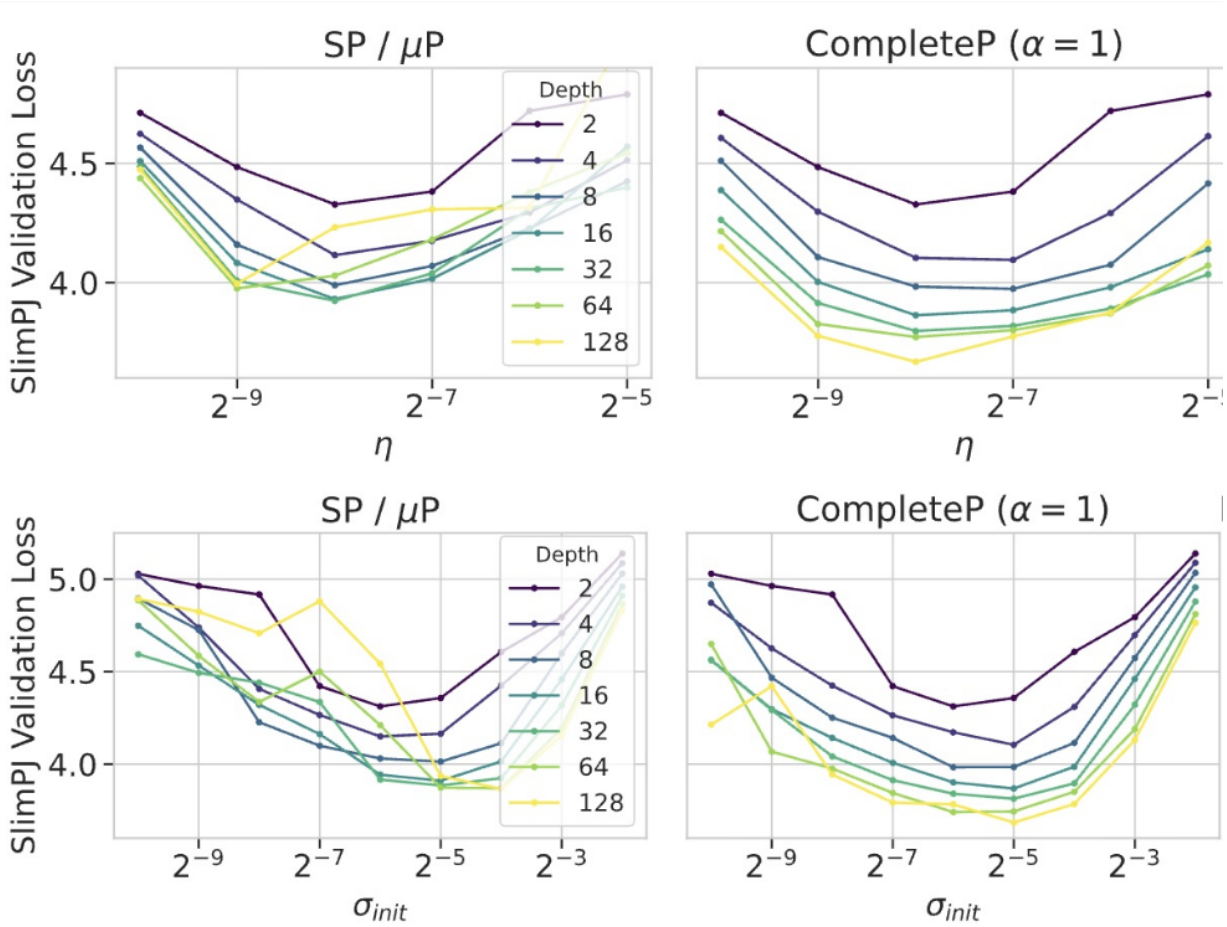
Infinite-width 2LP

High-dimensional 2LP

Deep ResNets

Large L, M, D limit

Conclusion



From "CompleteP" paper [Cerebras AI et al.,'25]. Transformer of varying depth trained on 300M tokens – here CompleteP = critical scale.

Large-scale limit in the MLU regime

Infinite-width 2LP

High-dimensional 2LP

Deep ResNets

Large L, M, D limit

Conclusion

	Tensor Program/DMFT	**Neural Mean ODE**
Approach	$M \propto D \to \infty$ then $L \to \infty$	$M, L \to \infty$ with D fixed
Other limitations	qualitative, 2nd step heuristic	loose upper bound
Main refs	[Yang et al., Bordelon et al.,...]	[Lu et al., Ding et al.,...]

Nb of params.	**8B**	**70B**	**405B**
L	32	80	126
D	4,096	8,192	16,384
M	14,336	28,672	53,248
M/D	3.5	3.5	3.25
ML/D	112	280	410

Shape hyperparameters of Llama 3.1 (MLP blocks)

$\hookrightarrow$ Unified theory for joint limits $L, M, D \to +\infty$? Quantitative rigorous rates?

Infinite-width 2LP

High-dimensional 2LP

Deep ResNets

Large L, M, D limit

Conclusion

1 Infinite-width 2LP
- Phase diagram
- Limit dynamics

2 High-dimensional 2LP
- Scalings
- Limit dynamics
- Cavity method

3 Deep ResNets
- Backward pass
- Stochastic approximation

4 Large L, M, D limit
- Scalings
- Phase diagram

5 Conclusion

Infinite-width 2LP
Phase diagram
Limit dynamics
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Conclusion

Two-Layer Perceptrons in the large-width limit

Definition and parameters

Infinite-width 2LP
Phase diagram
Limit dynamics
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Conclusion

Two-layer perceptrons (2LP):

$$f_\theta : \begin{cases} \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}, \\ x \mapsto \frac{1}{M} \sum_{i=1}^{M} v_i \rho(u_i^\top x). \end{cases}$$

Parameters $\theta = (u_i, v_i)_{i=1}^M \in (\mathbb{R}^{d_{\text{in}}} \times \mathbb{R})^M$.

Training set $(x_i, y_i)_{i=1}^n$, objective:

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \text{loss}(y_i, f_\theta(x_i)).$$

$\hookrightarrow$ After k GD steps on $\mathcal{L}$, what is the behaviour as $M \to +\infty$?

$$\begin{cases} u_0^j \sim \mathcal{N}(0, \sigma_u^2 \text{Id}_{d_{\text{in}}}), \\ v_0^j \sim \mathcal{N}(0, \sigma_v^2), \end{cases} \qquad \begin{cases} u_{k+1}^j = u_k^j - \eta_u M \nabla_{u^j} \mathcal{L}(\theta_k), \\ v_{k+1}^j = v_k^j - \eta_v M \nabla_{v^j} \mathcal{L}(\theta_k). \end{cases}$$

1. Phase diagram: scaling criteria

Infinite-width 2LP
Phase diagram
Limit dynamics
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Conclusion

$\hookrightarrow$ How to scale $\sigma_u, \sigma_v, \eta_u, \eta_v$ as $M \to +\infty$?

- Pre-activations: $P_k^j(x) = (u_k^j)^\top x$.
- Activations/Features: $A_k^j(x) = \rho(P_k^j(x))$.

Criterion 1 – Feature Diversity: $P_0^j(x) = \Theta(1)$

$$\mathbb{E}[P_0^j(x)] = 0, \qquad \mathbb{E}[P_0^j(x)^2] = \sigma_u^2 \|x\|_2^2 = \Theta(1).$$

$\hookrightarrow \sigma_u \sim 1/\|x\|_2 \sim 1.$

Criterion 2 – Non-explosion: $f_0(x) = \frac{1}{M} \sum_j v_0^j A_0^j = \mathcal{O}(1)$

$$\mathbb{E}[f_0(x)] = 0, \qquad \mathbb{E}[f_0(x)^2] = \frac{1}{M^2} \sum_j \sigma_v^2 \cdot \mathcal{O}(1) = \frac{\sigma_v^2}{M}.$$

$\hookrightarrow \sigma_v = \mathcal{O}(\sqrt{M})$.

2. Phase diagram: GD dynamics

Infinite-width 2LP
Phase diagram
Limit dynamics
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Conclusion

Let $g_t = \partial_2 \text{loss}(y, f_{\theta_t}(x))$. Continuous-time GD for $n = 1$:

$$\begin{cases} \dot{v}_t^j = -\eta_v \cdot M \cdot \nabla_{v^j} \mathcal{L}(\theta_t) = -\eta_v \cdot M \cdot \frac{1}{M} g_t \cdot \rho(P_t^j), \\ \dot{u}_t^j = -\eta_u \cdot M \cdot \nabla_{u^j} \mathcal{L}(\theta_t) = -\eta_u \cdot M \cdot \frac{1}{M} g_t \cdot v_t^j \cdot \rho'(P_t^j) x. \end{cases}$$

Evolution of pre-activations, predictor and loss:

$$\dot{P}_t^j = x^\top(\dot{u}_t^j) = -\eta_u \cdot v_t^j \cdot g_t \cdot \rho'(P_t^j) \|x\|_2^2,$$

$$\dot{f}_t = -\left[\frac{\eta_v}{M} \sum_j \rho(P_t^j)^2 + \frac{\eta_u}{M} \sum_j \rho'(P_t^j)^2 \|x\|_2^2 (v_t^j)^2 \right] g_t,$$

$$\dot{\mathcal{L}}_t = \dot{f}_t \cdot g_t = -\eta_v \cdot M \cdot \|\nabla_v \mathcal{L}(\theta_t)\|_2^2 - \eta_u \cdot M \cdot \|\nabla_u \mathcal{L}(\theta_t)\|_2^2$$

$$= -\frac{\eta_v}{M} \sum_j g_t^2 \cdot \rho^2(P_t^j) - \frac{\eta_u}{M} \sum_j g_t^2 \cdot \rho'^2(P_t^j) \cdot \|x\|_2^2 (V_t^j)^2$$

$$\sim \quad \eta_v \quad + \quad \eta_u \cdot (\sigma_v^2 + t^2 \eta_v^2).$$

3. Phase diagram: loss decay

Infinite-width 2LP
Phase diagram
Limit dynamics
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Conclusion

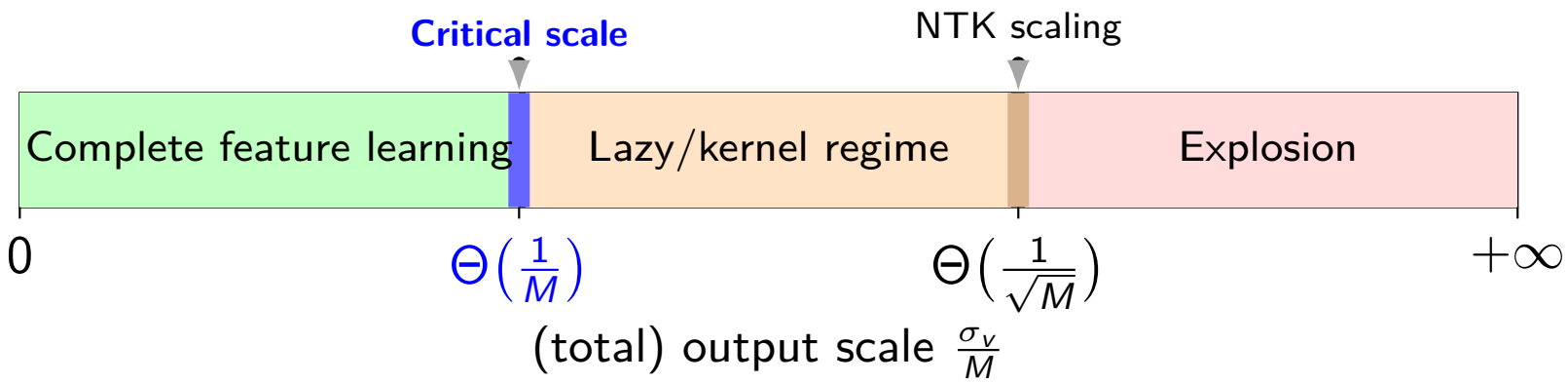
Criterion 3 – Balanced loss decay: $\dot{\mathcal{L}}_t = \Theta(1)$

$\hookrightarrow \eta_v \sim 1, \qquad \eta_u \sim \min(1, 1/\sigma_v^2).$

Criterion 4 – Feature learning: under criteria 1, 2, 3,

$$\dot{P}_t^j \sim \eta_u \cdot V_t^j \sim \eta_u(\sigma_v^2 + t^2\eta_v^2)^{1/2} \sim_{t=1} \begin{cases} 1/\sigma_v & \text{if } \sigma_v \geq 1, \\ 1 & \text{if } \sigma_v < 1. \end{cases}$$

Phase diagram:



[Dey et al.'25]

4. Phase diagram: operator viewpoint

Infinite-width 2LP
Phase diagram
Limit dynamics
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Conclusion

- $\alpha = 1/M$ can be absorbed into σ_v by adjusting η_v.
- The effective HPs are the effective scale $\alpha\sigma_v$ and LR $\alpha^2\eta$.

For a vector output $f_\theta(x) = \frac{1}{M} V\rho(Ux)$ with $V \in \mathbb{R}^{d_{\text{out}} \times M}$, a perturbation δa leads to

$$\|\delta f\|_{\text{RMS}} = \left\| \frac{1}{M} V \delta a \right\|_{\text{RMS}} \leq \frac{1}{M} \|V\|_{\text{RMS}\to\text{RMS}} \cdot \|\delta a\|_{\text{RMS}}$$

Heuristics using the operator norm of random matrices,

$$\|V_0\|_{\text{RMS}\to\text{RMS}} = \|V_0\|_{2\to 2} \cdot \frac{\sqrt{M}}{\sqrt{d_{\text{out}}}} \simeq \sigma_v(\sqrt{d_{\text{out}}} + \sqrt{M}) \cdot \frac{\sqrt{M}}{\sqrt{d_{\text{out}}}}$$

$$\implies \|\delta f\|_{\text{RMS}} \leq \|\delta a\|_{\text{RMS}} \cdot \sigma_v \cdot \left(\frac{1}{\sqrt{M}} + \frac{1}{\sqrt{d_{\text{out}}}} \right).$$

$\hookrightarrow$ Critical scale $\sigma_v \simeq \min(\sqrt{M}, \sqrt{d_{\text{out}}})$.

Limit dynamics in feature learning regime

Infinite-width 2LP
Phase diagram
Limit dynamics
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Conclusion

Neural network

$$f_\theta(x) = \frac{1}{M}\sum_{j=1}^{M} \phi(z^j, x), \qquad \begin{cases} \phi(z,x) = v\rho(u^\top x), \\ z^j = (u^j, v^j) \in \mathbb{R}^{d_{\text{in}}+d_{\text{out}}} =: \mathbb{R}^p. \end{cases}$$

GD step

$$\begin{cases} \hat{Z}^j_{k+1} = \hat{Z}^j_k - \frac{\eta}{n}\sum_{i=1}^n \nabla_z \phi(\hat{Z}^j_k, x_i)^\top \nabla_2 \text{loss}(y_i, \hat{f}_k(x_i)), \\ \hat{Z}^j_0 \sim_{\text{i.i.d.}} \mu_0 \in \mathcal{P}(\mathbb{R}^p). \end{cases}$$

Limit Dynamics: on some $(\Omega, \mathcal{F}, \mathbb{P})$, let $Z \in L^2(\Omega; \mathbb{R}^p)$ and

$$f_Z(x) := \mathbb{E}[\phi(Z, x)], \qquad \mathcal{L}(Z) = \frac{1}{n}\sum_{i=1}^n \text{loss}(y_i, f_Z(x_i)).$$

GD in L^2

$$\begin{cases} Z_{k+1} = Z_k - \frac{\eta}{n}\sum_{i=1}^n \nabla_z \phi(Z_k, x_i)^\top \nabla_2 \text{loss}(y_i, \hat{f}_k(x_i)), \\ Z_0 \sim \mu_0. \end{cases}$$

Quantitative limit theorem

Infinite-width 2LP
Phase diagram
Limit dynamics
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Conclusion

Theorem

If $\mathrm{Update}(Z_k, f_k)$ *is bounded-Lipschitz, then for every* $\delta \in (0,1]$,

$$\forall k \in \mathbb{N}, \exists c_k > 0 : \quad \max_{1 \le i \le n} |\hat{f}_k(x_i) - f_k(x_i)| \le c_k \sqrt{\log\left(\frac{n}{\delta}\right) \frac{D}{M}},$$

with proba $\ge 1 - \delta$.

- Closed dynamics for $\mathrm{Law}(Z_k)$ – Wasserstein gradient flow.
  In 2018: [Chizat,Bach], [Mei et al.], [Rotskoff et al.], [Sirignano et al.]...
- In the 2LP-case, the bound becomes

$$\ldots \le c_1 \log(n/\delta) \sqrt{D/M}, \quad \text{provided RHS } \le c_2.$$

- Converges to the limit we described iff $D/M \to 0$, which does not hold for deep ResNets architecture.
- Local convergence of the limit
  [Chizat,Colombo,Colombo,Fernandez-Real'26]

Proof sketch

Infinite-width 2LP
Phase diagram
Limit dynamics
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Conclusion

Let $(Z_k^j)_{k\geq 1}$ be i.i.d. copies of $(Z_k)_{k\geq 1}$ coupled such that $Z_0^j = \hat{Z}_0^j$. Define $\Delta_k = \max_j \|Z_k^j - \hat{Z}_k^j\|$. Since Update is Lipschitz,

$$\Delta_{k+1} \leq \Delta_k + c\big[\Delta_k + \max_i |\hat{f}_k(x_i) - f_k(x_i)|\big].$$

For each i,

$$|\hat{f}_k(x_i) - f_k(x_i)| \leq \left|\frac{1}{M}\sum_j \phi(\hat{Z}_k^j, x_i) - \frac{1}{M}\sum_j \phi(Z_k^j, x_i)\right| + \left|\frac{1}{M}\sum_j \phi(Z_k^j, x_i) - \mathbb{E}[\phi(Z, x_i)]\right| \leq c\Delta_k + \xi_{k,i}.$$

Hoeffding's inequality + union bound:

$$\max_{1\leq i\leq n} |\xi_{k,i}| \leq c\sqrt{\log\left(\frac{n}{\delta}\right)\frac{D}{M}}, \qquad \text{with proba} \geq 1-\delta.$$

$\hookrightarrow$ Conclusion: discrete Grönwall's lemma.

Lazy regime and empirical NTK

Infinite-width 2LP
Phase diagram
Limit dynamics
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Conclusion

Lazy Regime for $\sigma_v \gg 1$. Dynamics of the output:

$$\frac{d}{dt}\hat{f}_t(x) = -\frac{1}{n}\sum_{i=1}^{n} \hat{K}_t(x, x_i)\nabla_2 \mathrm{loss}(y_i, \hat{f}_t(x_i)),$$

where $\hat{K}_t$ is the Empirical Neural Tangent Kernel (NTK):

$$\hat{K}_t(x, x') := \frac{1}{M}\sum_{j=1}^{M} \nabla_z \phi(\hat{Z}_t^j, x)\nabla_z \phi(\hat{Z}_t^j, x')^\top$$

As $M \to \infty$, since $\hat{P}_t^j - \hat{P}_0^j = o(1)$ and $V_t^j - V_0^j = o(1)$,

$$\hat{K}_t(x, x') \xrightarrow[M\to+\infty]{} K_0(x, x') := \mathbb{E}[\nabla_z \phi(\hat{Z}_0, x)\nabla_z \phi(\hat{Z}_0, x')^\top].$$

$\hookrightarrow$ Dynamics closed in the predictor.

Recap

Infinite-width 2LP
Phase diagram
Limit dynamics
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Conclusion

→ **Criteria for scaling NN:**

Non-explosion, feature diversity at initialisation, balanced loss decay, feature learning.

↪ Phase diagram.

→ **For 2LP:** Infinite width limit at rate $O(\sqrt{D/M})$.

↪ Mean-equation, Wasserstein GF.

↪ **Converges to the limit we described – in the feature learning regime – iff** $D/M \to 0$, which does not hold for deep ResNets architecture.

Infinite-width 2LP
Phase diagram
Limit dynamics
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Conclusion

High-dimensional 2LP

Embedding

Infinite-width 2LP
High-dimensional 2LP
Scalings
Limit dynamics
Cavity method
Deep ResNets
Large L, M, D limit
Conclusion

Two-layer perceptrons (2LP):

$$h^0(x) = W_{\text{in}} x, \quad h(x) = \frac{1}{M} \sum_{j=1}^{M} \rho(\langle u^j, h^0(x) \rangle_{\overline{D}}) v^j \quad \in \mathbb{R}^D.$$

Parameters $\theta = (u_j, v_j)_{j=1}^M \in (\mathbb{R}^D \times \mathbb{R}^D)^M$.

(Un)embedding matrices: $W_{\text{in}} \in \mathbb{R}^D \times \mathbb{R}^{d_{\text{in}}}$, $W_{\text{out}} \in \mathbb{R}^D \times \mathbb{R}^{d_{\text{out}}}$.

Predictor: $f_\theta(x) = \frac{1}{D} W_{\text{out}}^\top h_1(x) \in \mathbb{R}^{d_{\text{out}}}$.

Initialisation

$$\begin{cases} u_0^j \sim \mathcal{N}(0, \sigma_u^2 \text{Id}_D), \\ v_0^j \sim \mathcal{N}(0, \sigma_v^2 \text{Id}_D), \end{cases} \qquad \begin{cases} W_{\text{in}}^d \sim \mathcal{N}(0, \sigma_{\text{in}}^2 \text{Id}_{d_{\text{in}}}), \\ W_{\text{out}}^d \sim \mathcal{N}(0, \sigma_{\text{out}}^2 \text{Id}_{d_{\text{out}}}). \end{cases}$$

$\hookrightarrow$ After k GD steps on $\mathcal{L}$, behaviour as $M, D \to +\infty$, $D \ll M$?

Scaling wrt dimension D

Infinite-width 2LP
High-dimensional 2LP
Scalings
Limit dynamics
Cavity method
Deep ResNets
Large L, M, D limit
Conclusion

- Pre-activations: $P_k^j(x) = \langle u_k^j, h^0(x) \rangle_{\overline{D}} := \frac{1}{D} \sum_{d=1}^{D} u_k^{j,d} h^{0,d}(x)$.
- Activations/Features: $A_k^j(x) = \rho(P_k^j(x))$.

Criterion 1 – Non-explosion: $h(x) = \frac{1}{M} \sum_j v_0^j A_0^j = \mathcal{O}(1)$

$$\mathbb{E}[\| W_{\text{in}} x \|_{\overline{D}}] = \Theta(1), \qquad \mathbb{E}[h(x)^2] = \frac{1}{M^2} \sum_j \sigma_v^2 \cdot \mathcal{O}(1) = \frac{\sigma_v^2}{M}.$$

$\hookrightarrow \quad \sigma_{\text{in}} = \Theta(1)$ (convention), $\quad \sigma_v = \mathcal{O}(\sqrt{M}), \quad \sigma_{\text{out}} = \mathcal{O}(\sqrt{D})$.

Criterion 2 – Feature Diversity: $P_0^j(x) = \Theta(1)$

$$\mathbb{E}[P_0^j(x)] = 0, \qquad \mathbb{E}[P_0^j(x)^2] = \frac{1}{D^2} \sum_{j=1}^{D} \sigma_u^2 (x^d)^2 = \Theta(1).$$

$\hookrightarrow \sigma_u = \Theta(\sqrt{D})$.

Training GD dynamics

Infinite-width 2LP
High-dimensional 2LP
Scalings
Limit dynamics
Cavity method
Deep ResNets
Large L, M, D limit
Conclusion

Let $g_t = \partial_2 \mathrm{loss}(y, f_{\theta_t}(x))$. Continuous-time GD for $n = 1$:

$$\begin{cases} \dot{v}_t^j = -\eta_v \cdot D \cdot M \cdot \nabla_{v^j} \mathcal{L}(\theta_t) = -\eta_v \rho(P_t^j) \cdot W_{\mathrm{out}} g_t, \\ \dot{u}_t^j = -\eta_u \cdot D \cdot M \cdot \nabla_{u^j} \mathcal{L}(\theta_t) = -\eta_u \rho'(P_t^j) \langle v_t^j, W_{\mathrm{out}} g_t \rangle_D h^0. \end{cases}$$

Evolution of pre-activations, predictor and loss:

$$\dot{P}_t^j = \langle \dot{u}_t^j, h^0 \rangle_D = -\eta_u \cdot \rho'(P_t^j) \langle v_t^j, W_{\mathrm{out}} g_t \rangle_D \|h^0\|_D^2,$$

$$\dot{f}_t = -\frac{\eta_v}{DM} \sum_j \rho(P_t^j)^2 W_{\mathrm{out}}^\top W_{\mathrm{out}} g_t - \frac{\eta_u}{DM} \sum_j \rho'(P_t^j)^2 \|h^0\|_D^2 \langle v_t^j, W_{\mathrm{out}} g_t \rangle_D W_{\mathrm{out}}^\top v_t^j,$$

$$\dot{\mathcal{L}}_t = \dot{f}_t \cdot g_t$$

$$= -\frac{\eta_v}{M} \sum_j \rho^2(P_t^j) \|W_{\mathrm{out}} g_t\|_D^2 - \frac{\eta_u}{M} \sum_j \rho'^2(P_t^j) \cdot \|h^0\|_D^2 \cdot \langle v_t^j, W_{\mathrm{out}} g_t \rangle_D^2$$

$$\sim \quad \eta_v \quad + \quad \eta_u \cdot \frac{\sigma_v^2 + t^2 \eta_v^2}{D}.$$

Phase diagram wrt dimension

Infinite-width 2LP
High-dimensional 2LP
Scalings
Limit dynamics
Cavity method
Deep ResNets
Large L, M, D limit
Conclusion

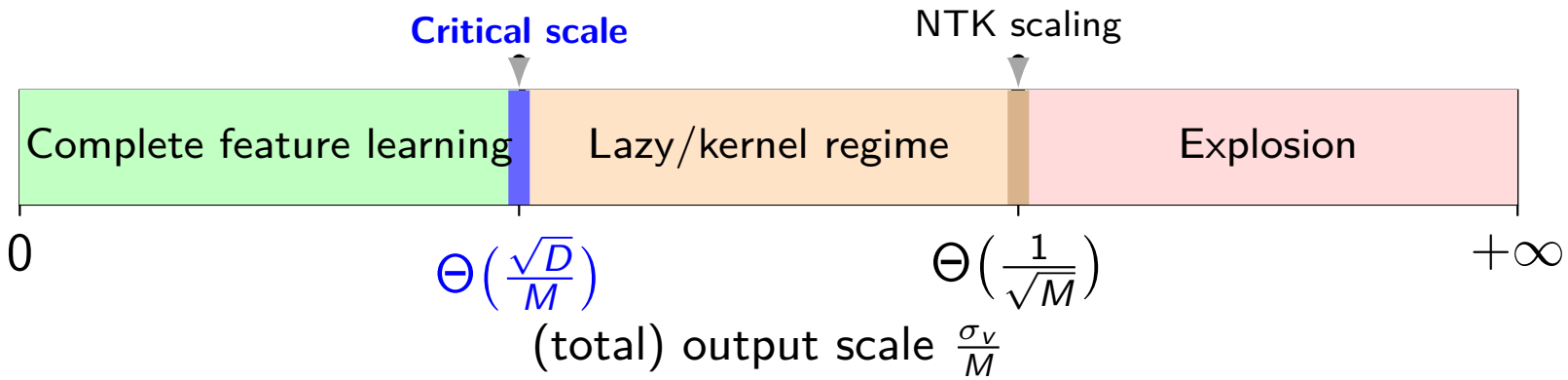
Criterion 3 – Balanced loss decay: $\dot{\mathcal{L}}_t = \Theta(1)$

$\hookrightarrow \eta_v = \mathcal{O}(1), \qquad \eta_u \eta_v^2 = \mathcal{O}(1), \qquad \eta_u = \mathcal{O}(D/\sigma_v^2).$

Criterion 4 – Feature learning: under criteria 1, 2, 3,

$$\dot{P}_t^j \sim \eta_u \cdot V_t^j \sim \eta_u (\sigma_v^2/D + t^2 \eta_v^2)^{1/2}, \qquad \begin{cases} \sigma_{\text{out}} = \Theta(1), \\ \eta_u = \eta_v = \Theta(1), \\ \sigma_v = \mathcal{O}(\sqrt{D}). \end{cases}$$

Phase diagram:



Limit dynamics

Infinite-width 2LP
High-dimensional 2LP
Scalings
Limit dynamics
Cavity method
Deep ResNets
Large L, M, D limit
Conclusion

Neural-mean equation

$$h^0(x) = W_{\text{in}} x, \quad h_k(x) = \mathbb{E}[\rho(\langle U_k, h^0(x)\rangle_{\overline{D}}) V_k | W_{\text{in}}, W_{\text{out}}] \quad \in \mathbb{R}^D.$$

Predictor: $\hat{f}_k(x) = \frac{1}{D} W_{\text{out}}^\top h_k(x) \in \mathbb{R}^{d_{\text{out}}}$.

Training dynamics for $n = 1$

$$\begin{cases} \Delta V_{k+1} = -\eta_v \rho(\langle U_k, h^0(x)\rangle_{\overline{D}}) W_{\text{out}} \partial_2 \mathcal{L}(y, f_k(x)), \\ \Delta U_{k+1} = -\eta_u \rho'(\langle U_k, h^0(x)\rangle_{\overline{D}}) \partial_2 \mathcal{L}(y, f_k(x)) \langle V_k, W_{\text{out}} \partial_2 \mathcal{L}(y, f_k(x))\rangle_{\overline{D}} h^0(x). \end{cases}$$

Theorem (Quantitative convergence)

$$\forall k \in \mathbb{N}, \exists c_k^1, c_k^2 > 0 : \quad \max_{1 \le i \le n} |\hat{f}_k(x_i) - f_k(x_i)| \le c_k^1 \sqrt{\log\left(\frac{n}{\delta}\right) \frac{D}{M}},$$

with proba $\ge 1 - \delta$, *provided that RHS* $\le c_k^2$.

Tracking correlations

Infinite-width 2LP
High-dimensional 2LP
Scalings
Limit dynamics
Cavity method
Deep ResNets
Large L, M, D limit
Conclusion

NEW GOAL: sending $D \to +\infty$.

For the d-th coordinate:

$$h_k^d(x) = \mathbb{E}[\rho(\langle U_k^{(D)}, h^{0,(D)}(x)\rangle_{\overline{D}}) V_k^d \,|\, W_{\text{in}}, W_{\text{out}}],$$

where

$$U_k^d = \sqrt{D} U^d + \Delta U_k^d, \qquad V_k^d = \sqrt{D} V^d + \Delta V_k^d,$$

for $U^d \sim \mathcal{N}(0, \sigma_u^2)$, $V^d \sim \mathcal{N}(0, \sigma_v^2)$. Consequently,

$$\langle U_k^{(D)}, h_k^{(D)}\rangle_{\overline{D}} = \underbrace{\frac{1}{\sqrt{D}} \sum_{d=1}^{D} U^d h_k^d}_{\text{CLT term}} + \underbrace{\frac{1}{D} \sum_{d=1}^{D} \Delta U_k^d h_k^d}_{\text{LLN term}}.$$

$\perp\!\!\!\perp V_k^d$ — **CORRELATED**

$\Rightarrow$ How to track correlations along the training?

Skeleton maps

Infinite-width 2LP
High-dimensional 2LP
Scalings
Limit dynamics
Cavity method
Deep ResNets
Large L, M, D limit
Conclusion

CLT-type sums

$$(\mathbf{S}^{\mathbf{h}}, \mathbf{S}^{\mathbf{b}}) := \left(\left(\frac{1}{\sqrt{D}} \sum_{d=1}^{D} U^d h_i^d(s), \frac{1}{\sqrt{D}} \sum_{d=1}^{D} V^d [W_{\text{out}} \nabla \mathcal{L}(f_i(x))]^d \right) \right)_{0 \le i \le K-1}$$

Skeleton maps

$$(\Delta U_k, \Delta V_k) = (F_k \cdot h^0(x), G_k \cdot W_{\text{out}} \nabla \mathcal{L}(f_k(x)))$$

$$F_k = F_k(\mathbf{S}^{\mathbf{h}}, \mathbf{S}^{\mathbf{b}}, \|h^0\|_D^2, \|W_{\text{out}} \nabla \mathcal{L}(f_k(x))\|_D^2).$$

Definition

Functions $\mathbf{F}, \mathbf{G} : \mathbb{R}^K \times \mathbb{R}^K \times \mathbb{R}^2 \to \mathbb{R}^K$ such that

$$\begin{cases} F_{k+1} = F_k - \eta_u \rho'(z_k^h + F_k \cdot \|h^0\|_D^2) \\ \qquad\qquad\qquad\qquad [z_k^b + F_k \cdot \|W_{\text{out}} \nabla \mathcal{L}(f_k(x))\|_D^2], \\ G_{k+1} = G_k - \eta_v \rho(z_k^h + F_k \cdot \|h^0\|_D^2), \end{cases}$$

evaluating at $(\mathbf{z}^h, \mathbf{z}^b, \|h^0\|_D^2, \|W_{\text{out}} \nabla \mathcal{L}(f_k(x))\|_D^2)$.

Cavity method

Infinite-width 2LP
High-dimensional 2LP
Scalings
Limit dynamics
Cavity method
Deep ResNets
Large L, M, D limit
Conclusion

Taylor expansion – recalling $\mathbf{S}^{\mathbf{b}} = (\frac{1}{\sqrt{D}} \sum_{d=1}^{D} V^d b_i^d)_{0 \le i \le K-1}$ –

$$
\begin{aligned}
&\mathbb{E}[\rho(\langle U_k^{(D)}, h^{0,(D)} \rangle_{\overline{D}}) \sqrt{D} V^d | W_{\text{in}}, W_{\text{out}}] \\
&= \mathbb{E}[\rho(S_k^{\mathbf{h}} + \langle F(\mathbf{S}^{\mathbf{h}}, \mathbf{S}^{\mathbf{b}}, \cdot), h^{0,(D)} \rangle_{\overline{D}}) \sqrt{D} V^d | W_{\text{in}}, W_{\text{out}}] \\
&\simeq \mathbb{E}[\rho(S_k^{\mathbf{h}} + \langle F(\mathbf{S}^{\mathbf{h}}, \overline{\mathbf{S}}^{\mathbf{b}}, \cdot), h^{0,(D)} \rangle_{\overline{D}}) \sqrt{D} V^d | W_{\text{in}}, W_{\text{out}}] \\
&+ \mathbb{E}[\rho'(S_k^{\mathbf{h}} + \langle F(\mathbf{S}^{\mathbf{h}}, \overline{\mathbf{S}}^{\mathbf{b}}, \cdot), h^{0,(D)} \rangle_{\overline{D}}) \langle \nabla_{z^b} F(\mathbf{S}^{\mathbf{h}}, \overline{\mathbf{S}}^{\mathbf{b}}, \cdot)[V^d \mathbf{b}^d], h^{0,(D)} \rangle_{\overline{D}} V^d] \\
&+ O(D^{-1/2}),
\end{aligned}
$$

where $\overline{\mathbf{S}}^{\mathbf{b}^D} := \mathbf{S}^{\mathbf{b}} - V^d \mathbf{b}^d / \sqrt{D}$, using

$$(\mathbf{S}^{\mathbf{h}}, \overline{\mathbf{S}}^{\mathbf{b}}) \perp\!\!\!\perp V^d \quad | \quad (W_{\text{in}}, W_{\text{out}}).$$

Quantitative CLT

$$\mathbb{E}[F(\mathbf{S}^{\mathbf{h}}, \overline{\mathbf{S}}^{\mathbf{b}}, \cdot) | W_{\text{in}}, W_{\text{out}}] = \mathbb{E}[F(\mathbf{Z}^{\mathbf{h}}, \mathbf{Z}^{\mathbf{b}}, \cdot) | W_{\text{in}}, W_{\text{out}}] + O(D^{-1/2}).$$

Quantitative convergence

Infinite-width 2LP
High-dimensional 2LP
Scalings
Limit dynamics
Cavity method
Deep ResNets
Large L, M, D limit
Conclusion

$\to$ As $D \to +\infty$, coordinates become independent:

$$\hat{f}_k(x) \xrightarrow[D\to+\infty]{} \bar{f}(x), \quad h_k^d \xrightarrow[D\to+\infty]{} H_k, \quad \|h^{0,(D)}\|_D^2 \xrightarrow[D\to+\infty]{} \mathbb{E}[H_0^2],$$
$$\|W_{\text{out}}^{(D)} \nabla \mathcal{L}(\hat{f}_k(x))\|_D^2 \xrightarrow[D\to+\infty]{} \mathbb{E}[W_{\text{out}} \nabla \mathcal{L}(\bar{f}_k(x))].$$

Limit dynamics

$$H_k = \sigma_v^2 \mathbb{E}[\rho'(P_k(s)) \mathbb{E}[H_0(s) \nabla_{z^b} F_k(\mathbf{Z}^h, \mathbf{Z}^b, \cdot) | \mathbf{Z}^h, \mathbf{Z}^b]] \cdot [W_{\text{out}} \nabla \mathcal{L}(\bar{f}_i(x))]_{i \le k-1}$$
$$+ \mathbb{E}[\rho(P_k(s)) G_k(s, \mathbf{Z}^h, \mathbf{Z}^b, \mathbb{E}[H_0^2], \mathbb{E}[W_{\text{out}} \nabla \mathcal{L}(\bar{f}_k(x))]) | W_{\text{in}}, W_{\text{out}}],$$
$$P_k(s) := Z_k^h(s) + \mathbb{E}[H_0(s) F_k(\mathbf{Z}^h, \mathbf{Z}^b, \cdot) | \mathbf{Z}^h, \mathbf{Z}^b].$$

Theorem

$$\forall \delta \in (0,1], \exists c_k, c_k' > 0 : \ \|h_k^{(D)} - H_k^{(D)}\|_D \le c_k \frac{1 + \log(1/\delta)}{\sqrt{D}},$$

proba $\ge 1 - \delta$, *provided RHS* $\le c_k'$.

Recap

Infinite-width 2LP
High-dimensional 2LP
Scalings
Limit dynamics
Cavity method
Deep ResNets
Large L, M, D limit
Conclusion

→ **Criteria for scaling NN:**

Non-explosion, feature diversity at initialisation, balanced loss decay, feature learning.

→ **For 2LP:**

- Infinite width limit at rate $\mathcal{O}(\sqrt{D/M})$.

  ↪ Mean-equation, Wasserstein GF.

- High-dimensional limit at rate $\mathcal{O}(1/\sqrt{D})$.

  ↪ Cavity method, history-dependent dynamics.

↪ **Converges with rate** $\mathcal{O}(1/L + D/M)$ when $D \ll M$, which does not hold for deep ResNets architecture.

↪ But this is still a useful toy model.

Infinite-width 2LP

High-dimensional 2LP

Scalings

Limit dynamics

Cavity method

Deep ResNets

Large L, M, D limit

Conclusion

ResNets in the large-depth limit

Definition and parameters

Infinite-width 2LP
High-dimensional 2LP
Deep ResNets
Backward pass
Stochastic approximation
Large L, M, D limit
Conclusion

Residual Network with initial state $h_\theta^0(x) = x \in \mathbb{R}^D$:

$$h_\theta^\ell(x) = h_\theta^{\ell-1}(x) + \frac{1}{LM} \sum_{j=1}^{M} \phi(h_\theta^{\ell-1}(x), z^{j,\ell}), \quad \ell \in [\![1, L]\!]$$

Parameters $\theta = (z^{j,\ell})_{j,\ell} \in (\mathbb{R}^p)^{L \times M}$.

Example (2LP blocks): $\phi(h, z) = v\rho(u^\top x)$.

Training set $(x_i, y_i)_{i=1}^n$, objective:

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{loss}(y_i, h_\theta^L(x_i)).$$

$\hookrightarrow$ After k GD steps on $\mathcal{L}$, what is the behaviour as $L, M \to +\infty$?

$\hookrightarrow$ Discretised ODE!

Backward pass and training Dynamics

Infinite-width 2LP
High-dimensional 2LP
Deep ResNets
Backward pass
Stochastic approximation
Large L, M, D limit
Conclusion

Chain rule

$$\frac{\partial \mathcal{L}}{\partial z^{j,\ell}} = \frac{\partial \mathcal{L}}{\partial h^{\ell}} \cdot \frac{\partial h^{\ell}}{\partial z^{j,\ell}} = \frac{1}{LM} \nabla_z \phi(h_\theta^{\ell-1}, z^{j,\ell})^\top b^{\ell}.$$

The backward pass

$$b^{\ell} := \left[\frac{\partial h_k^L}{\partial h_k^{\ell}}\right]^\top \nabla \mathcal{L}(f_k(x))$$

satisfies the adjoint equation

$$\begin{cases} b^{\ell-1} = b^{\ell} + \frac{1}{LM} \sum_{j=1}^{M} [\nabla_h \phi(h^{\ell-1}, z^{j,\ell})]^\top b^{\ell}, \\ b^{L} = \nabla_h \mathrm{loss}(y, h^L). \end{cases}$$

Training dynamics:

$$\hat{Z}_{k+1}^{j,\ell} = \hat{Z}_k^{j,\ell} - \eta [\nabla_z \phi(\hat{h}_k^{\ell-1}, \hat{Z}_k^{j,\ell})]^\top \hat{b}_k^{\ell}.$$

$\to$ Limit $L, M \to +\infty$ for fixed $s = \ell/L \in [0, 1]$.

Quantitative Limit

Infinite-width 2LP
High-dimensional 2LP
Deep ResNets
Backward pass
Stochastic approximation
Large L, M, D limit
Conclusion

Limit dynamics: the Mean ODE

$$\begin{cases} \partial_s h_k(s) = \mathbb{E}[\phi(h_k(s), Z_k(s))], \\ h_k(0) = x, \\ \partial_s b_k(s) = -\mathbb{E}[\nabla_h \phi(h_k(s), Z_k(s))^\top b_k(s)], \\ b_k(1) = \nabla_2 \mathrm{loss}(y, h_k(1)). \end{cases}$$

Initialising LM copies $Z_0^{j,\ell}(s) \equiv \hat{Z}_0^{j,\ell}$, let $\Delta_k^Z := \sup_{j,\ell} \|Z_k^{j,\ell}(s_l) - \hat{Z}_k^{j,\ell}\|$.

Theorem (Chizat'25)

If $\mathrm{Update}(Z_k^\ell, h_k^\ell(1))$ *is bounded-Lipschitz, then for every* $\delta \in (0, 1]$,

$$\forall k \in \mathbb{N}, \exists c_k > 0 : \max\left\{\Delta_k^z, \Delta_k^h, \Delta_k^b\right\} \leq c_k \left[\frac{1}{L} + \sqrt{\frac{1 + \log(1/\delta)}{LM}}\right],$$

with proba $\geq 1 - \delta$.

- This bound is empirically tight for small k.
- $L \to +\infty$ is enough to ensure convergence for fixed M.

Stochastic approximation for Mean ODEs

Infinite-width 2LP
High-dimensional 2LP
Deep ResNets
Backward pass
Stochastic approximation
Large L, M, D limit
Conclusion

Consider the general mean ODE

$$\partial_s a(s) = \mathbb{E}[f(s, a(s), Z(s))],$$

for $z \mapsto Z(s)$ Lipschitz and f bounded-Lipschitz.
Integration using inexact Euler-Monte Carlo method

$$\hat{a}^\ell = \hat{a}^{\ell-1} + \frac{1}{LM} \sum_{j=1}^{M} \hat{f}((\ell-1)/L, \hat{a}^{\ell-1}, \hat{Z}^{j,\ell}),$$

Let $(Z^{j,\ell})_{j,\ell}$ be i.i.d. samples such that $\|\hat{Z}^{j,\ell} - Z^{j,\ell}(\ell/L)\| \leq \varepsilon$.

Proposition

For every $\delta \in (0, 1/2]$, with proba $\geq 1 - \delta$,

$$\sup_\ell \|\hat{a}^\ell - a(s_\ell)\| \leq \|\hat{a}^0 - a(0)\| + \|f - \hat{f}\|_\infty + \varepsilon + \frac{1}{L} + \sqrt{\frac{\log(1/\delta)}{M}}.$$

$\hookrightarrow$ Applied to $f_k^h(s, a, z) = \phi(z, a)$, $f_k^b(s, a, z) = D_z\phi(z, h_k(s))^\top a$ + propagation of Lipschitz-regularity for $s \mapsto Z(s)$.

Proof sketch

Infinite-width 2LP
High-dimensional 2LP
Deep ResNets
Backward pass
Stochastic approximation
Large L, M, D limit
Conclusion

Grönwall lemma from $\Delta_{\ell+1} \leq \Delta_\ell + \mathcal{E}^{\ell+1}_{\text{Euler}} + \mathcal{E}^{\ell+1}_{\text{MC}} + \mathcal{E}^{\ell+1}_{\text{approx}}$, where:

- $\sum_\ell \mathcal{E}^{\ell+1}_{\text{Euler}}$ (discretization):

$$\sum_\ell \int_{s_\ell}^{s_{\ell+1}} \mathbb{E}[f(s, a(s), Z(s))] - \mathbb{E}[f(s_\ell, a(s_\ell), Z(s_\ell))]\mathrm{d}s = \mathcal{O}\left(\frac{1}{L}\right).$$

- $\sum_\ell \mathcal{E}^{\ell+1}_{\text{MC}}$ (Monte-Carlo fluctuation):

$$\frac{1}{LM}\sum_{j,\ell} \mathbb{E}[f(s_\ell, a(s_\ell), Z(s_\ell))] - f(s_\ell, a(s_\ell), Z^{j,\ell}(s_\ell)) = \mathcal{O}\left(\sqrt{\frac{\log(L/\delta)}{LM}}\right),$$

via Hoeffding / martingale argument to control $\sup_\ell$.

- $\sum_\ell \mathcal{E}^{\ell+1}_{\text{approx}}$ (approximation):

$$\frac{1}{LM}\sum_{j,\ell} f(s_\ell, a(s_\ell), Z^{j,\ell}(s_\ell)) - \hat{f}(s_\ell, \hat{a}^\ell, \hat{Z}^{j,\ell}) \leq \frac{1}{L}[\varepsilon + \|f - \hat{f}\|_\infty + \Delta_\ell].$$

Recap

Infinite-width 2LP
High-dimensional 2LP
Deep ResNets
Backward pass
Stochastic approximation
Large L, M, D limit
Conclusion

For Residual Neural Networks:

- Limit dynamics: the Neural Mean ODE.

  $\hookrightarrow$ Stochastic approximation lemma.

- ODE discretisation error $\mathcal{O}(1/L)$.

- Monte-Carlo error $\mathcal{O}(1/\sqrt{LM})$.

$\hookrightarrow$ Deep architecture converge to the feature learning limit even when $D \propto M$!

Infinite-width 2LP

High-dimensional 2LP

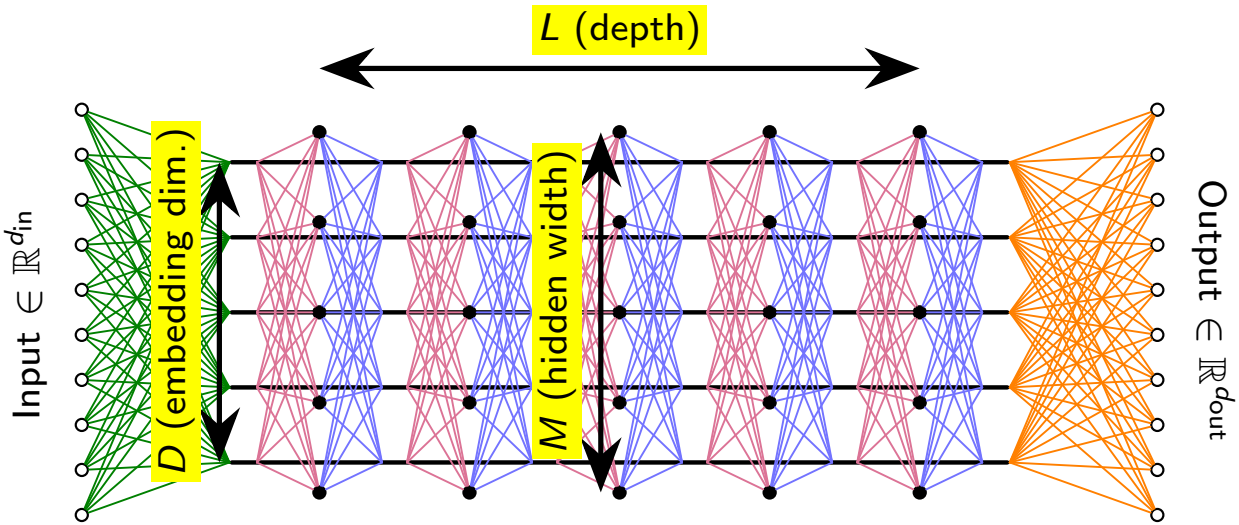
Deep ResNets

Large L, M, D limit

Scalings

Phase diagram

Conclusion



ResNets in the large L, M, D limit

Definition and parameters

Infinite-width 2LP
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Scalings
Phase diagram
Conclusion

Residual Network with 2LP blocks and embedding

$$\begin{cases} \hat{h}^0(x) &= W_{\text{in}} x, \\ \hat{h}^\ell(x) &= \hat{h}^{\ell-1}(x) + \frac{1}{LM} \sum_{j=1}^{M} v^{j,\ell} \rho(\langle u^{j,\ell}, \hat{h}^{\ell-1}(x) \rangle_{\overline{D}}), \quad \ell \in [\![1, L]\!], \\ \hat{f}(x) &= \frac{1}{D} W_{\text{out}}^T \hat{h}^L(x). \end{cases}$$

Parameters θ

$$W_{\text{in}} \in \mathbb{R}^{D \times d_{\text{in}}}, \quad W_{\text{out}} \in \mathbb{R}^{D \times d_{\text{out}}}, \quad (u^{j,\ell}, v^{j,\ell}) \in (\mathbb{R}^D \times \mathbb{R}^D)^{M \times L}.$$

Initialisation

$$\begin{cases} W_{\text{in}}[i,j] \sim_{\text{i.i.d}} \mathcal{N}(0, \sigma_{\text{in}}^2), \\ W_{\text{out}}[i,j] \sim_{\text{i.i.d}} \mathcal{N}(0, \sigma_{\text{out}}^2), \end{cases} \qquad \begin{cases} u_0^{j,\ell}, u_i^{j,\ell} \sim_{\text{i.i.d}} \mathcal{N}(0, \sigma_u^2), \\ v_0^{j,\ell}, v_i^{j,\ell} \sim_{\text{i.i.d}} \mathcal{N}(0, \sigma_v^2). \end{cases}$$

GD Updates:

$$\begin{cases} u_{k+1}^{j,\ell} &= u_k^{j,\ell} - \eta_u \cdot L \cdot M \cdot D \cdot \nabla_{u^{j,\ell}} \mathcal{L}(\theta_k), \\ v_{k+1}^{j,\ell} &= v_k^{j,\ell} - \eta_v \cdot L \cdot M \cdot D \cdot \nabla_{v^{j,\ell}} \mathcal{L}(\theta_k). \end{cases}$$

Scaling criteria

Infinite-width 2LP
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Scalings
Phase diagram
Conclusion

Criterion 1 – Non explosion: $\|h^\ell\|_{\overline{D}} = \Theta(1), \quad \|f\|_{\overline{D}} = \mathcal{O}(1).$

$\hookrightarrow \sigma_{\text{in}} = \Theta(1/\|x\|_2) \sim 1$ (convention), $\quad \sigma_{\text{out}} = \mathcal{O}(\sqrt{D})$.

Criterion 2 – Feature diversity: $\mathbb{E}(P_0^{j,\ell})^2 = \Theta(1)$

$\hookrightarrow$ Under Criterion 1, $\quad \sigma_u = \Theta(\sqrt{D})$.

$\rightarrow$ Signal propagation at initialisation:

$$h_0^\ell = h_0^0 + \frac{1}{LM} \sum_{\ell'=1}^{\ell} \sum_{j=1}^{M} v_0^{j,\ell} \rho(\langle u_0^{j,\ell}, h_0^{\ell-1} \rangle_{\overline{D}}),$$

so that $\mathbb{E}[h_0^\ell] = \mathbb{E}[h_0^0] = 0$, and conditionally on W_{in}, and

$$\mathbb{E}\|h_0^\ell\|_{\overline{D}}^2 = \mathbb{E}\|h_0^\ell - h_0^0 + h_0^0\|_{\overline{D}}^2 = \mathbb{E}\|h^\ell - h_0^0\|_{\overline{D}}^2 + \mathbb{E}\|h_0^0\|_{\overline{D}}^2. \quad (*)$$

Signal propagation at initialisation

Infinite-width 2LP
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Scalings
Phase diagram
Conclusion

Let $\Delta_\ell^2 := \mathbb{E}\|h^\ell - h^0\|_D^2$ and $\mathcal{F}_\ell := \sigma(W_{\text{in}}, Z^{j,1}, \dots, Z^{\ell,1})$.

$$\mathbb{E}[\|h^{\ell+1} - h^0\|_D^2 \mid \mathcal{F}_\ell] = \|h^\ell - h^0\|_D^2 + \frac{2}{LM}\sum_{j=1}^M \mathbb{E}[\langle h^\ell - h^0, v_0^{j,\ell+1}\rho(\langle u_0^{j,\ell+1}, h^\ell\rangle_D)\rangle_D \mid \mathcal{F}_\ell] + \left(\frac{1}{LM}\right)^2 \mathbb{E}\left[\left\|\sum_{j=1}^M v_0^{j,\ell+1}\rho(\langle u_s^{j,\ell+1}, h^\ell\rangle_D)\right\|_D^2 \,\middle|\, \mathcal{F}_\ell\right].$$

For simplicity, if $|\rho(x)| < c|x|$, then

$$\mathbb{E}[\|h^{\ell+1} - h^0\|_D^2 \mid \mathcal{F}_\ell] = \|h^\ell - h^0\|_D^2 + \frac{M\cdot\sigma_v^2}{(LM)^2}\cdot c\left(\frac{\sigma_u}{\sqrt{D}}\cdot\|h^\ell\|_D\right)^2 = \|h^\ell - h^0\|_D^2 + \frac{c\sigma_v^2}{L^2M}\|h^\ell\|_D^2. \quad (\sigma_u \sim \sqrt{D})$$

Hence, taking $\mathbb{E}$ and using $(*)$,

$$\Delta_{\ell+1}^2 \le \left(1 + \frac{c\sigma_v^2}{L^2M}\right)\Delta_\ell^2 + \frac{c\sigma_v^2}{L^2M}\|h^0\|_D^2.$$

Learning Criterions

Infinite-width 2LP
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Scalings
Phase diagram
Conclusion

Integrating,

$$\Delta_\ell^2 \leq \left(\left(1+\frac{c\sigma_v^2}{L^2M}\right)^\ell - 1\right)\|h^0\|_D^2 = \left[\exp\left(\frac{c\sigma_v^2\ell}{L^2M}(1+o(1))\right)-1\right]\cdot\|h^0\|_D^2.$$

Hence,

$$\Delta_L = \begin{cases} \Theta\left(\frac{\sigma_v^2}{LM}\right), & \text{if } \sigma_v = \mathcal{O}(\sqrt{LM}), \\ \text{explodes otherwise.} \end{cases}$$

Criterion 3 – Balanced local decay:
$(U^{j,\ell})$ and $(V^{j,\ell})$ contribute to loss decay in $\Theta(1)$.

Criterion 4 – Feature learning:
$\mathbb{E}[|\partial_t P^{j,\ell}|^2] = \Theta(1) \Rightarrow \sigma_{\text{out}} = \Theta(1)$.

Criterion 5 – Complete feature learning: Maximal Local Update.

Computing updates

Infinite-width 2LP
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Scalings
Phase diagram
Conclusion

→ Gradient flow equations: let $b^\ell := D \cdot \left[\frac{\partial \mathcal{L}}{\partial h^\ell}\right]^\top$ denote the rescaled backward pass, so that $\|b^\ell\|_D \sim 1$.

$$\begin{cases} \partial_t u_t^{j,\ell} = -\eta_u \rho'(P_t^{j,\ell}) \cdot Q_t^{j,\ell} \cdot b_t^{j,\ell} \cdot h_t^{\ell-1}, \\ \partial_t v_t^{j,\ell} = -\eta_v \cdot \rho(P_t^{j,\ell}) \cdot b_t^\ell, \end{cases} \qquad \begin{cases} P_t^{j,\ell} := \langle u_t^{j,\ell}, h_t^{\ell-1} \rangle_D, \\ Q_t^{j,\ell} := \langle v_t^{j,\ell}, b_t^{\ell-1} \rangle_D. \end{cases}$$

For small t: $\quad \mathbb{E}(P_t^{j,\ell})^2 \sim 1, \quad \mathbb{E}(Q_t^{j,\ell})^2 \sim \frac{\sigma_v^2}{D} + t^2 \eta_v^2.$

→ Pre-activation updates:

$$\partial_t P_t^{j,\ell} = \underbrace{\langle \partial_t v^{j,\ell}, h_t^{\ell-1} \rangle_D}_{\text{local update}} + \underbrace{\langle u_t^{j,\ell}, \partial_t h_t^{\ell-1} \rangle_D}_{\text{global update}}.$$

$$\partial_t^{\text{loc}} P_t^{j,\ell} = \langle \partial_t u_t^{j,\ell}, h_t^{\ell-1} \rangle_D = -\eta_u \cdot \rho'(P_t^{j,\ell}) \cdot Q_t^{j,\ell} \cdot \|h_t^{\ell-1}\|_D.$$

→ Loss decay:

$$\partial_t \mathcal{L}_t = \underbrace{-\frac{\eta_u}{LM} \cdot \sum_{j,\ell} \rho'(P_t^{j,\ell})^2 \cdot (Q_t^{j,\ell})^2 \cdot \|h_t^{\ell-1}\|_D^2}_{\eta_u \cdot (\sigma_v^2/D + t^2 \eta_v^2)} \underbrace{- \frac{\eta_v}{LM} \cdot \sum_{j \neq \ell} \rho(P_t^{j,\ell})^2 \cdot \|b_t^\ell\|_2^2}_{\eta_v}.$$

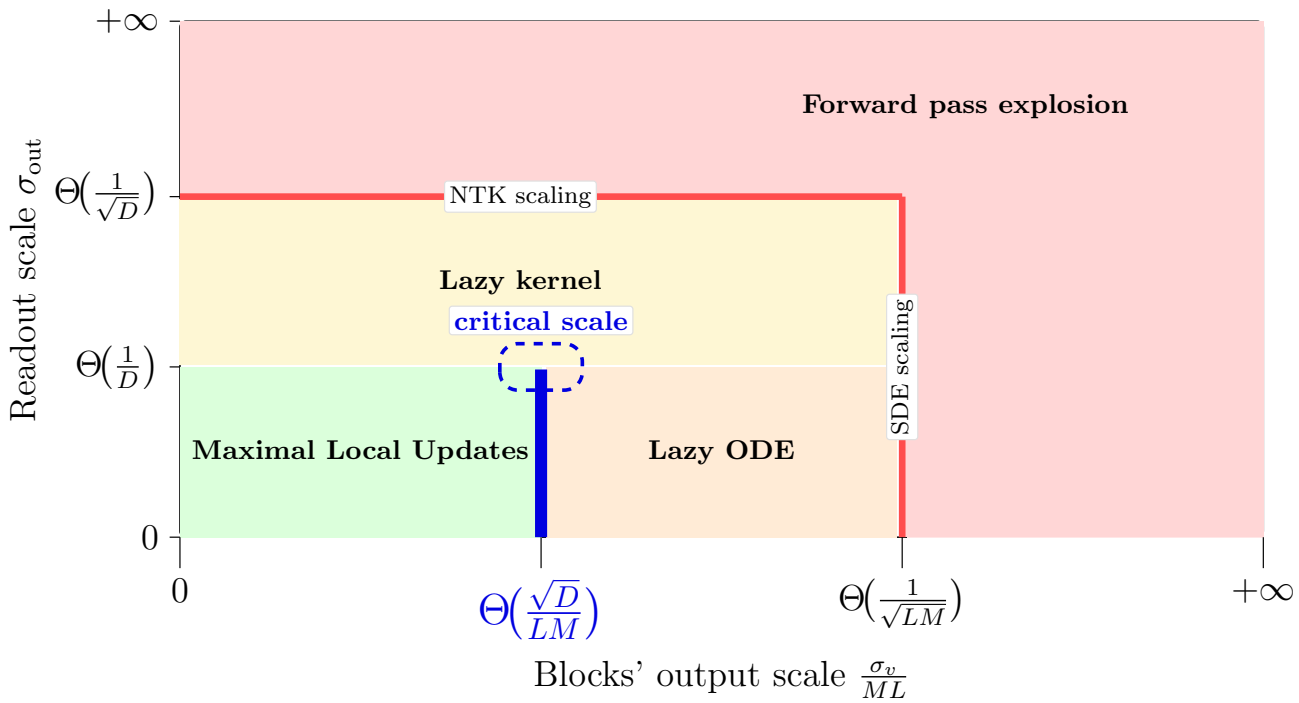
Phase diagram

Infinite-width 2LP
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Scalings
Phase diagram
Conclusion

Criteria 3 – Loss decay: $\eta_v \sim 1, \quad \eta_u \sim \min(D/\sigma_v^2, 1)$.

Criteria 5 – MLU:

$$\mathbb{E}|\partial_t^{\text{loc}} P_t^{j,\ell}|^2 \sim \eta_u \cdot \left(\frac{\sigma_v^2}{D} + t^2\eta_v^2\right)^{1/2} \sim \begin{cases} 1/\sigma_v & \text{if } \sigma_v \geq \sqrt{D}, \\ 1 & \text{if } \sigma_v < \sqrt{D}. \end{cases}$$



[Chizat et al.'18], [Yang et al.'21], [Dey et al.'25], [Chizat'25]...

Conclusion

Infinite-width 2LP

High-dimensional 2LP

Deep ResNets

Large L, M, D limit

Conclusion

→ **Criteria for scaling NN:**

Non-explosion, feature diversity at initialisation, balanced loss decay, feature learning.

→ **For 2LP:**

- Infinite width limit at rate $\mathcal{O}(\sqrt{D/M})$.

  ↪ Mean-equation, Wasserstein GF.

- High-dimensional limit at rate $\mathcal{O}(1/\sqrt{D})$.

  ↪ Cavity method, history-dependent dynamics.

→ **For Deep ResNets:** convergence allows for $M \propto D$,

and we will prove the $\mathcal{O}\left(1/L + \sqrt{D/(LM)} + 1/\sqrt{D}\right)$ rate.

Large Depth ResNet with D Dependence

Infinite-width 2LP

High-dimensional 2LP

Deep ResNets

Large L, M, D limit

Conclusion

Theorem (informal)

Assume the Feature Learning regime and gradient clipping hold.

$$\max_{0\leq \ell \leq L} \|\hat{h}_k^\ell - H_k^{(D)}(s_\ell)\|_{\overline{D}} \leq c\left(\frac{1}{L} + \frac{\sqrt{D} + \log(1/\delta)}{\sqrt{LM}} + \frac{1}{\sqrt{D}}\right).$$

Optimal Scaling Under Budget Constraint

Fix total number of parameters $P = D(d_{\text{in}} + d_{\text{out}}) + 2MLD$. The error bound is minimized by selecting the scaling dimensions:

$$D \asymp P^{1/3}, \quad M \cdot L \asymp P^{2/3}, \quad L \geq P^{1/6}$$

Conjecture: $\|\hat{f}_k - f_k\|_{\overline{D}} = \mathcal{O}\left(\frac{1}{L} + \frac{D}{ML} + \frac{1}{\sqrt{D}}\right)$.

This leads to the Optimal Shape: $L \sim P^{1/5}$, $D \sim P^{2/5}$, $M \sim P^{2/5}$.

Infinite-width 2LP
High-dimensional 2LP
Deep ResNets
Large L, M, D limit
Conclusion

Proof Intuition: Cavity Method (DMFT)

Decompose the coordinates of the hidden states h_k^d:

$$U_k^d(s) = \sqrt{D}U^d + \Delta U_k^d(s),$$

$$\langle U_k^{(D)}(s), h_k^{(D)}(s)\rangle_{\overline{D}} = \underbrace{\frac{1}{\sqrt{D}}\sum_{d=1}^{D} U_0^d h_k^d(s)}_{\text{CLT term}} + \underbrace{\frac{1}{D}\sum_{d=1}^{D} \Delta U_k^d h_k^d(s)}_{\text{LLN term}}.$$

Open Directions

- Unified theory for all joint limits $L, M, D \to \infty$.
- Complete phase diagram tracking SDE limits vs local limits?
- Learning properties: understanding mathematically why this specific architecture is so special.
- Tightening theoretical connections to empirical practice.