ResNets of All Shapes and Sizes

Louis-Pierre Chaintron (EPFL)

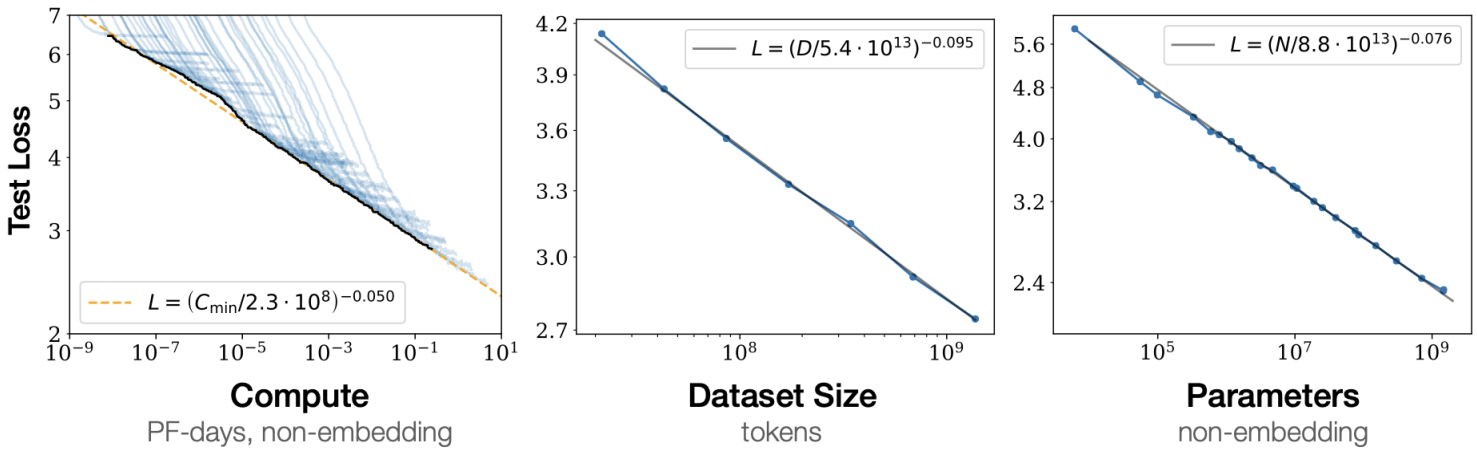
ProbAI Theory of Scaling Laws Workshop 2026
– University of Warwick, June 23rd

Joint work with Lénaïc Chizat and Javier Maass (EPFL).

Scaling up Neural Networks

Observation:

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

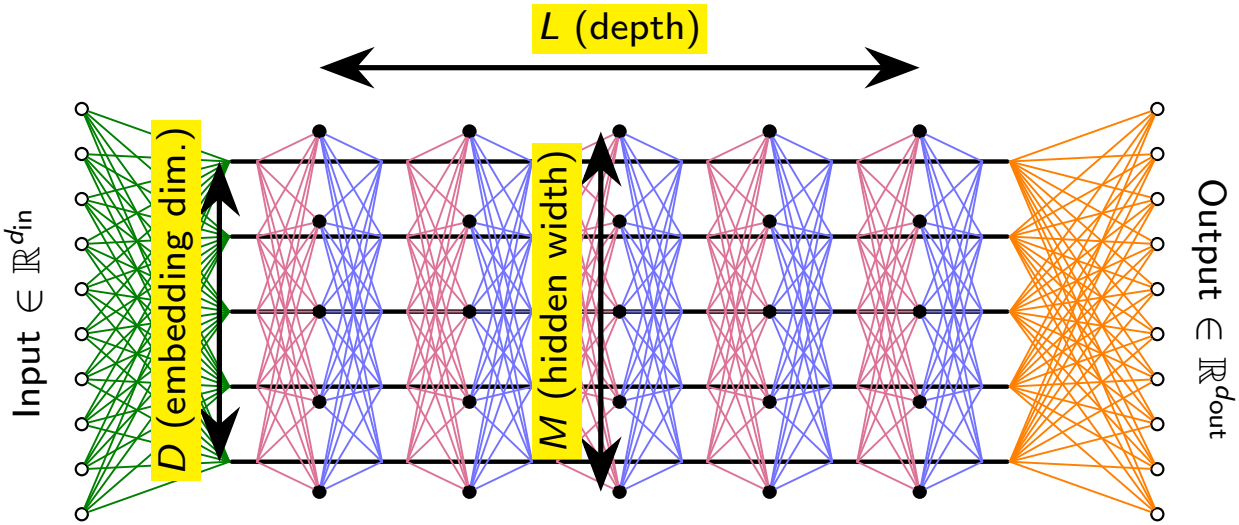


Performance vs compute [Kaplan et al'20]
N parameters, dataset size D, compute C

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

$\Rightarrow$ Improve choices of HP.

Residual Neural Networks



$$\begin{cases} h^0(x) = W_{\text{in}}\, x \in \mathbb{R}^D, \\ \qquad \text{input} \qquad\qquad \text{activation } \rho : \mathbb{R} \to \mathbb{R} \\ h^\ell(x) = h^{\ell-1}(x) + \frac{1}{LM} \sum_{j=1}^{M} v^{j,\ell}\, \rho\big(u^{j,\ell} \cdot h^{\ell-1}(x)\big), \ \ell \in [1:L], \\ \qquad\qquad\qquad\qquad\qquad \text{pre-activation} \\ y(x) = W_{\text{out}}^\top h^L(x) \in \mathbb{R}^{d_{\text{out}}}. \end{cases}$$

Parameters θ:

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

Training dynamics

- Training set $(x_i)_{1\le i\le n}$, loss function:

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Loss}_i(y(x_i)).$$

- Gradient descent:

$$\begin{cases} W_0^{\mathrm{in}}[i,j] \overset{iid}{\sim} \mathcal{SG}(0,\sigma_{\mathrm{in}}^2), \\ u_0^{j,\ell}[i] \overset{iid}{\sim} \mathcal{SG}(0,\sigma_u^2), \\ v_0^{j,\ell}[i] \overset{iid}{\sim} \mathcal{SG}(0,\sigma_v^2), \\ W_0^{\mathrm{out}}[i,j] \overset{iid}{\sim} \mathcal{SG}(0,\sigma_{\mathrm{out}}^2), \end{cases} \qquad \begin{cases} W_{k+1}^{\mathrm{in}} = W_k^{\mathrm{in}} - \eta_{\mathrm{in}} \nabla_{W^{\mathrm{in}}} \mathcal{L}(\theta_k), \\ u_{k+1}^{j,\ell} = u_k^{j,\ell} - \eta_u \nabla_{u^{j,\ell}} \mathcal{L}(\theta_k), \\ v_{k+1}^{j,\ell} = v_k^{j,\ell} - \eta_v \nabla_{v^{j,\ell}} \mathcal{L}(\theta_k), \\ W_{k+1}^{\mathrm{out}} = W_k^{\mathrm{out}} - \eta_{\mathrm{out}} \nabla_{W^{\mathrm{out}}} \mathcal{L}(\theta_k). \end{cases}$$

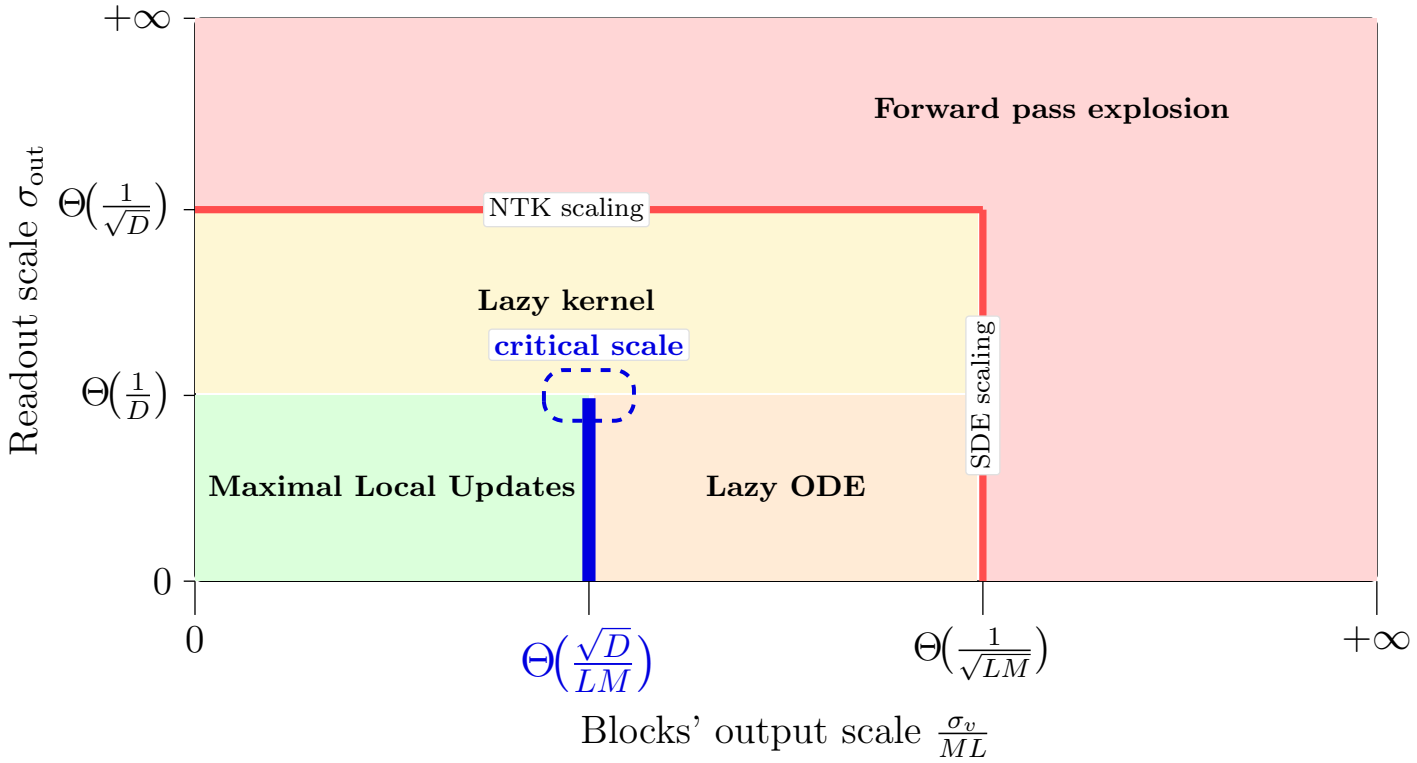
Given $(\sigma_{\mathrm{in}}, \sigma_u, \sigma_v, \sigma_{\mathrm{out}})$, we scale $(\eta_{\mathrm{in}}, \eta_u, \eta_v, \eta_{\mathrm{out}})$ so that:

- Loss decay due to W_{in}, W_{out} is $\Theta(1)$.
- Loss decay due to $u^{j,\ell}$, $v^{j,\ell}$ is $\Theta(1/(LM))$.

Pre-activation in $\Theta(1)$: we choose $\sigma_{\mathrm{in}} \sim 1$ and $\sigma_u = \sigma_v \sim 1/\sqrt{D}$.

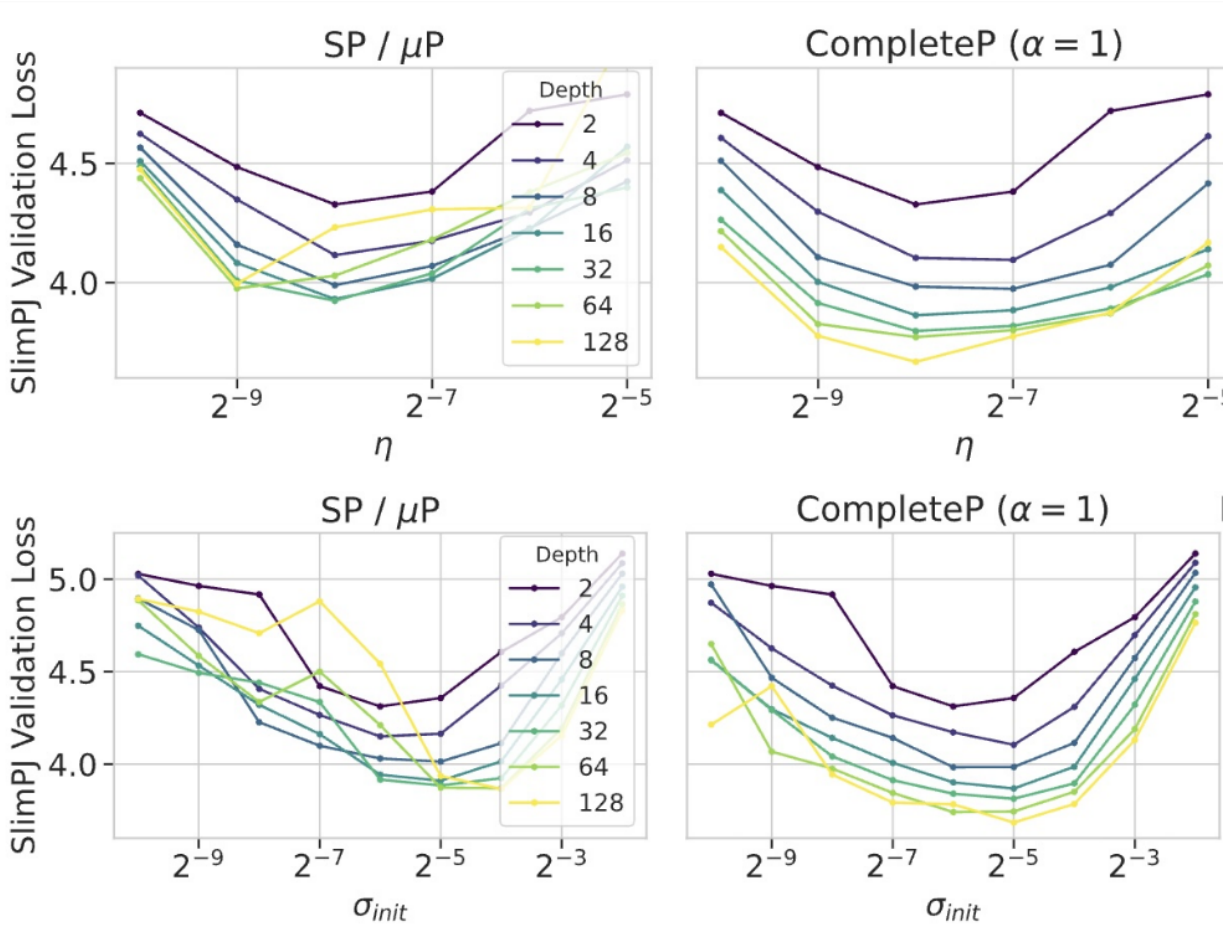
Phase diagram (assuming $ML \gg D \gg 1$)

- Lazy kernel: (pre)activations update in $o(1)$ at any step.
- Feature learning: (pre)activations update in $\Theta(1)$ in the first GD step.
- Maximal Local Update (MLU): contribution in $\Theta(1)$ of local weights to update.
- Lazy ODE: feature learning but not local – local contributions in $o(1)$.



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

Application: HP transfer [Yang et al'22]



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

	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?

Scaling limits

Limit theorem

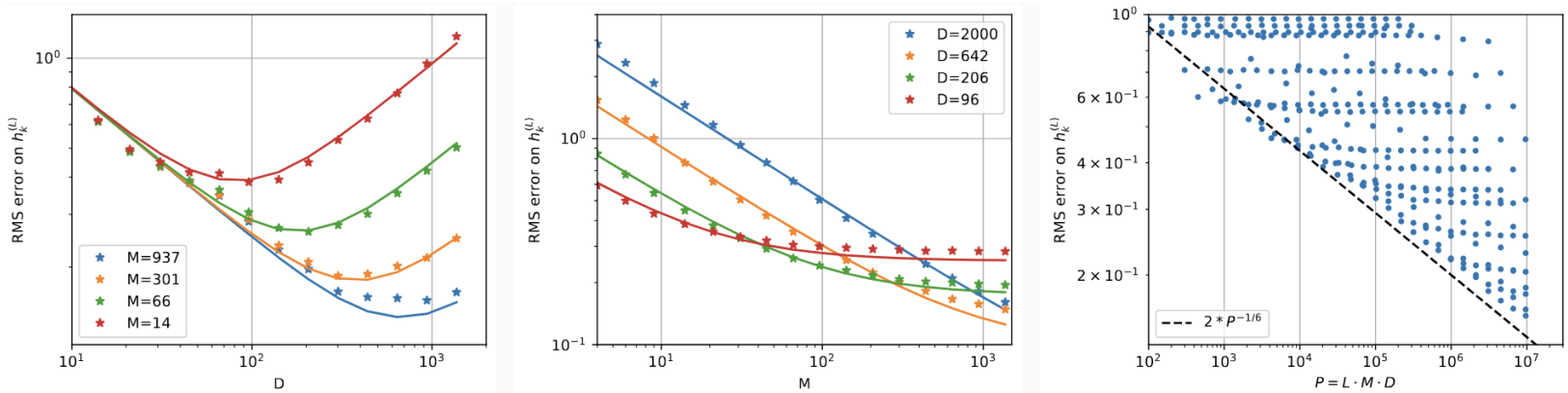
Theorem (C., Chizat, Maass'26)

$\exists$ *(limit dynamics)* : $\forall k \geq 1,\ \exists c_k, c'_k > 0 :\ \forall \delta \in (0,1]$,

$$(error)_k \leq c_k \left[\frac{1}{L} + \sqrt{\frac{D}{LM}} + \frac{1+\log(1/\delta)}{\sqrt{D}} \right] \quad \textit{with proba} \geq 1-\delta,$$

provided RHS $\leq c'_k$ *– for clipped GD.*

$\hookrightarrow$ For a budget $P = LMD$, we get $(\text{error}) = O(P^{-1/6})$.



Error in embedding space for fitted coefficients at $k = 15$.

Backward pass

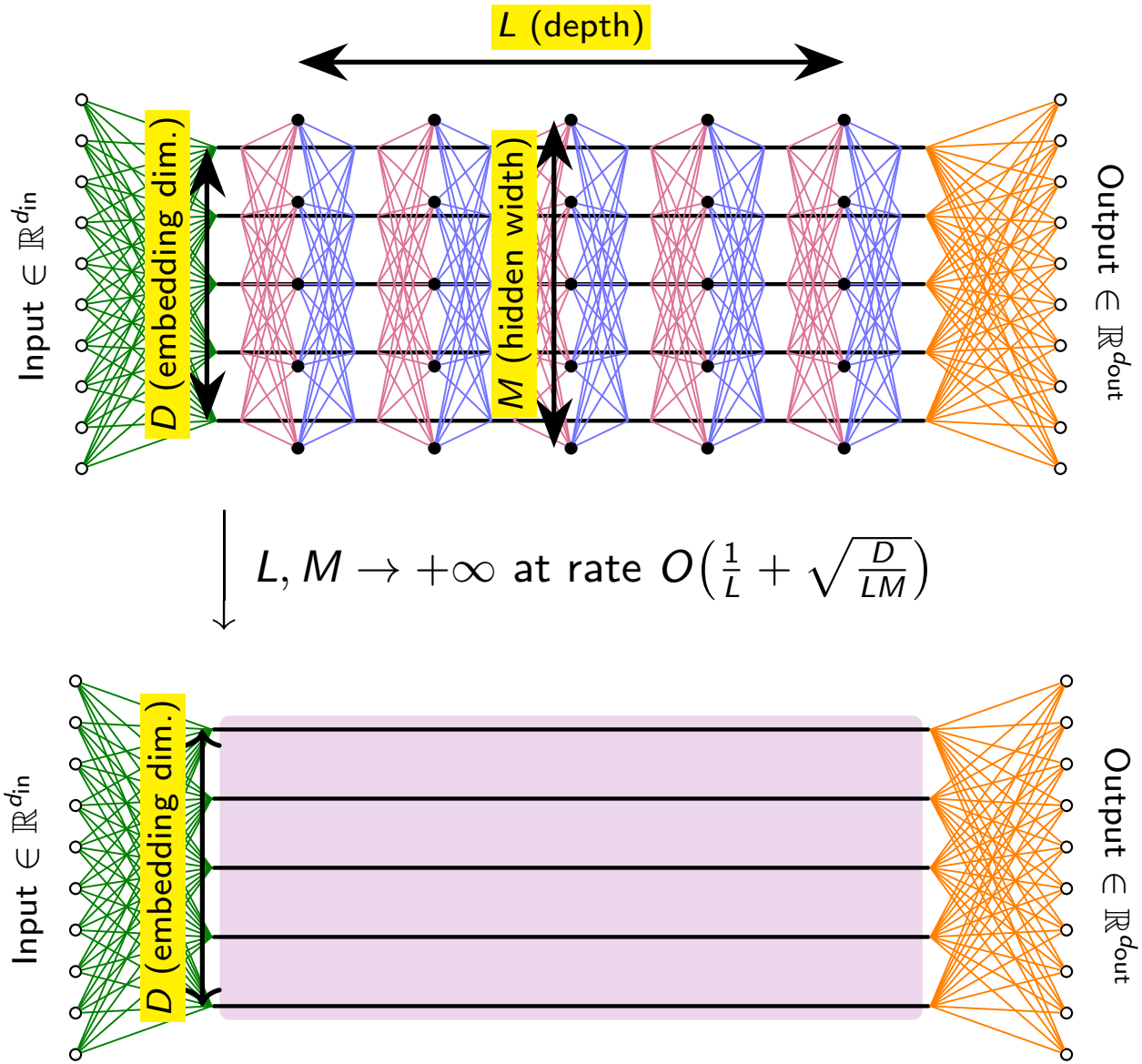
Training dynamics:

$$\begin{cases} h_k^0 = W_{\text{in}} x, \\ h_k^\ell = h_k^{\ell-1} + \frac{1}{LM} \sum_{j=1}^M \rho(\langle u_k^{j,\ell}, h_k^{\ell-1} \rangle_{\overline{D}}) v_k^{j,\ell}, \\ f_k(x) = W_{\text{out}}^\top h_k^L, \\ b_k^{\ell-1} = b_k^\ell + \frac{1}{LM} \sum_{j=1}^M \rho'(\langle u_k^{j,\ell}, h_k^{\ell-1} \rangle_{\overline{D}}) \underline{\langle v_k^{j,\ell}, b_k^{\ell-1} \rangle_{\overline{D}}} u_k^{j,\ell}, \\ b_k^L = W_{\text{out}}^\top \nabla \mathcal{L}(f_k(x)), \\ u_{k+1}^{j,\ell} = u_k^{j,\ell} - \eta_u \rho'(\langle u_k^{j,\ell}, h_k^{\ell-1} \rangle_{\overline{D}}) \underline{\langle v_k^{j,\ell}, b_k^{\ell-1} \rangle_{\overline{D}}} h_k^{\ell-1}, \\ v_{k+1}^{j,\ell} = v_k^{j,\ell} - \eta_v \rho(\langle u_k^{j,\ell}, h_k^{\ell-1} \rangle_{\overline{D}}) b_k^{\ell-1}. \end{cases}$$

Normalized backward pass:

$$b_k^\ell := D W_{\text{out}} \left[\frac{\partial h_k^L}{\partial h_k^\ell} \right]^\top \nabla \mathcal{L}(f_k(x)), \qquad \begin{cases} u_0^{j,\ell} \sim \mathcal{SG}(0, D\sigma_u^2), \\ v_0^{j,\ell} \sim \mathcal{SG}(0, D\sigma_v^2). \end{cases}$$

Neural Mean ODE



Neural Mean ODE

Limit $L, M \to +\infty$ dynamics:

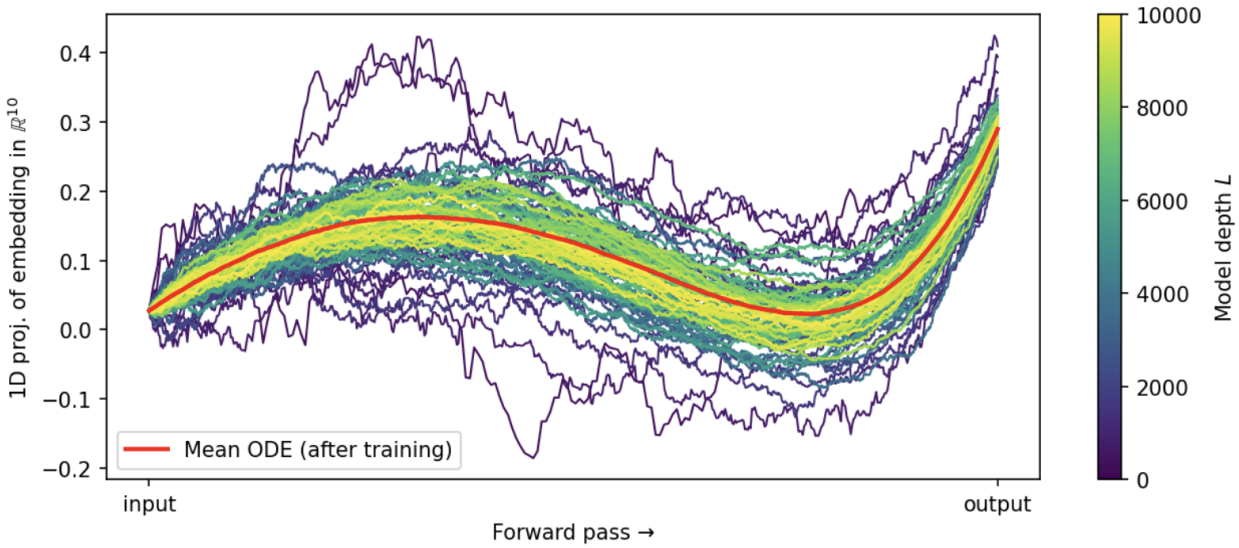
$$\begin{cases} \partial_s h_k(s) = \mathbb{E}[V_k(s)\rho(\langle U_k(s), h_k(s)\rangle_{\overline{D}})], \\ \partial_s b_k(s) = -\mathbb{E}[V_k(s)\rho'(\langle U_k(s), h_k(s)\rangle_{\overline{D}})\underline{\langle V_k(s), b_k(s)\rangle_{\overline{D}}} U_k(s)], \\ U_{k+1}(s) = U_k(s) - \eta_u \rho'(\langle U_k(s), h_k(s)\rangle_{\overline{D}})\underline{\langle V_k(s), b_k(s)\rangle_{\overline{D}}} h_k(s), \\ V_{k+1}(s) = V_k(s) - \eta_v \rho(\langle U_k(s), h_k(s)\rangle_{\overline{D}}) b_k(s). \end{cases}$$

Theorem (Chizat'25)

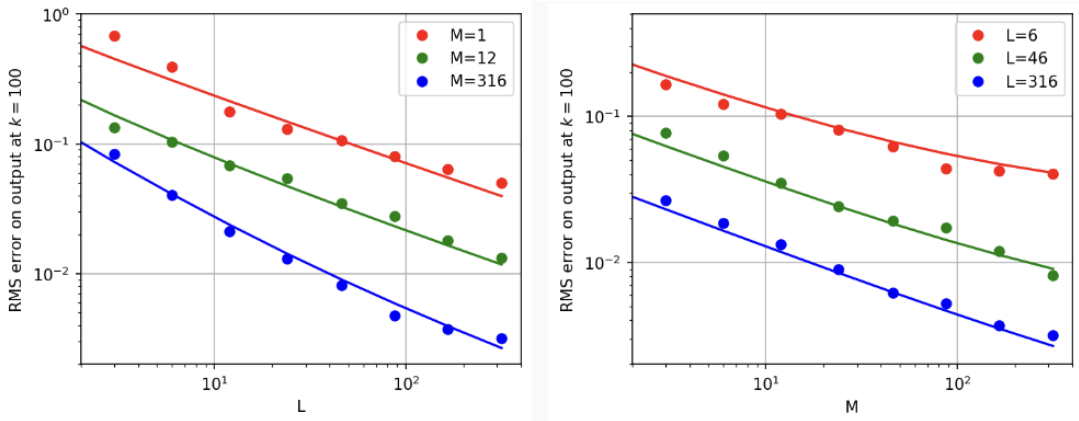
For Clipped GD in the MLU regime, if $\log L \vee D \lesssim M$, *then with high probability*

$$\max_{0 \le \ell \le L} |h_k^\ell(x) - h_k(\ell/L, x)| = O\left(\frac{1}{L} + \sqrt{\frac{D}{LM}}\right).$$

Experiment: narrow vs deep ResNet



1D projection of the forward pass on one sample after 50 GD steps – GD for the square-loss on n $= 10$ training samples with $D = 10$, $M = 1$.



Experiment vs theory (with fitted coefficients) after k $= 100$ GD steps.

Large D limit

Neural mean ODE in $\mathbb{R}^D$:

$$\begin{cases} \partial_s h_k(s) = \mathbb{E}[V_k(s)\rho(\langle U_k(s), h_k(s)\rangle_{\overline{D}})], \\ \partial_s b_k(s) = -\mathbb{E}[V_k(s)\rho'(\langle U_k(s), h_k(s)\rangle_{\overline{D}})\langle V_k(s), b_k(s)\rangle_{\overline{D}} U_k(s)], \\ U_{k+1}(s) = U_k(s) - \eta_u \rho'(\langle U_k(s), h_k(s)\rangle_{\overline{D}})\langle V_k(s), b_k(s)\rangle_{\overline{D}} h_k(s), \\ V_{k+1}(s) = V_k(s) - \eta_v \rho(\langle U_k(s), h_k(s)\rangle_{\overline{D}}) b_k(s). \end{cases}$$

Initialisation:

$$U_k(s) = \underbrace{U_0(s)} + \underbrace{\Delta U_k(s)} = \sqrt{D}U + \Delta U_k(s),$$

$$\mathcal{SG}(0, D\sigma_u^2) \quad O(1)$$

so that

$$\langle U_k(s), h_k(s)\rangle_{\overline{D}} = \frac{1}{\sqrt{D}} \sum_{d=1}^{D} U^d h_k^d(s) + \frac{1}{D} \sum_{d=1}^{D} \Delta U_k^d(s) h_k^d(s),$$

for $(U^d)_{d\geq 1}$ i.i.d. ~ 1.

Skeleton maps

CLT-type sums

$$(\mathbf{S}^{\mathbf{h}^D}(s), \mathbf{S}^{\mathbf{b}^D}(s)) := \left(\left(\frac{1}{\sqrt{D}} \sum_{d=1}^{D} U^d h_i^d(s), \frac{1}{\sqrt{D}} \sum_{d=1}^{D} V^d b_i^d(s) \right) \right)_{0 \le i \le K-1}$$

Skeleton maps

$$(\Delta U_k^{(D)}(s), \Delta V_k^{(D)}(s)) = (f_k^{\mathbf{h}^D, \mathbf{b}^D}, g_k^{\mathbf{h}^D, \mathbf{b}^D})(s, \mathbf{S}^{\mathbf{h}^D}(s), \mathbf{S}^{\mathbf{b}^D}(s))$$

Definition

Functions $\mathbf{f}^{\mathbf{h}^D, \mathbf{b}^D}, \mathbf{g}^{\mathbf{h}^D, \mathbf{b}^D} : [0,1] \times \mathbb{R}^K \times \mathbb{R}^K \to (\mathbb{R}^D)^K$ such that

$$\begin{cases} f_{k+1}^{\mathbf{h}^D, \mathbf{b}^D} = f_k^{\mathbf{h}^D, \mathbf{b}^D} - \eta_u \rho'(z_k^h + \langle h_k^D, f_k^{\mathbf{h}^D, \mathbf{b}^D} \rangle_{\overline{D}})[z_k^b + \langle b_k^D, g_k^{\mathbf{h}^D, \mathbf{b}^D} \rangle_{\overline{D}}] h_k^D, \\ g_{k+1}^{\mathbf{h}^D, \mathbf{b}^D} = g_k^{\mathbf{h}^D, \mathbf{b}^D} - \eta_v \rho(z_k^h + \langle h_k^D, f_k^{\mathbf{h}^D, \mathbf{b}^D} \rangle_{\overline{D}}) b_k^D, \end{cases}$$

evaluating at $(s, \mathbf{z}^h, \mathbf{z}^b)$.

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

Limit skeleton

Definition (Mean-field Skeleton)

Random maps $\mathbf{F}^{\mathbf{H},\mathbf{B}}, \mathbf{G}^{\mathbf{H},\mathbf{B}} : [0,1] \times \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}^K$ given by

$$\begin{cases} F_{k+1}^{\mathbf{H},\mathbf{B}} = F_k^{\mathbf{H},\mathbf{B}} - \eta_u \rho'(z_k^h + \mathbb{E}[H_k F_k^{\mathbf{H},\mathbf{B}}])[z_k^b + \mathbb{E}[B_k G_k^{\mathbf{H},\mathbf{B}}]]H_k, \\ G_{k+1}^{\mathbf{H},\mathbf{B}} = G_k^{\mathbf{H},\mathbf{B}} - \eta_v \rho(z_k^h + \mathbb{E}[H_k F_k^{\mathbf{H},\mathbf{B}}])B_k, \end{cases}$$

The finite-D dynamics, $1 \leq d \leq D$,

$$\begin{aligned} \partial_s h_k^d(s) &= \mathbb{E}[V_k^d(s)\rho(\langle U_k(s), h_k(s)\rangle_{\overline{D}})] \\ &= \mathbb{E}[(\sqrt{D}V^d + (g_k^{\mathbf{h}^D,\mathbf{b}^D})^d)\rho(\langle \sqrt{D}U_k + f_k^{\mathbf{h}^D,\mathbf{b}^D}, h_k(s)\rangle_{\overline{D}})], \end{aligned}$$

becomes, as $D \to +\infty$,

$$\begin{aligned} \partial_s H_k(s) = \sigma_v^2 \mathbb{E}[\rho'(P_k(s))\mathbb{E}[H_k(s)\nabla_{z^b} F_k^{\mathbf{H},\mathbf{B}}(s, \mathbf{Z}^h, \mathbf{Z}^b)|\mathbf{Z}^h, \mathbf{Z}^b]]\cdot \mathbf{B}_{\wedge k-1}(s) \\ + \mathbb{E}[\rho(P_k(s))G_k^{\mathbf{H},\mathbf{B}}(s, \mathbf{Z}^h, \mathbf{Z}^b)|W_{\text{in}}, W_{\text{out}}], \end{aligned}$$

$$P_k(s) := Z_k^h(s) + \mathbb{E}[H_k(s)F_k^{\mathbf{H},\mathbf{B}}(s, \mathbf{Z}^h, \mathbf{Z}^b)|\mathbf{Z}^h, \mathbf{Z}^b].$$

Limit system

Forward pass:

$$\begin{cases} H_k(0) = W_{\text{in}} \cdot x, \\ P_k(s) := Z_k^h(s) + \mathbb{E}[H_k(s) F_k^{\mathbf{H},\mathbf{B}}(s, \mathbf{Z}^h, \mathbf{Z}^b) | \mathbf{Z}^h, \mathbf{Z}^b], \\ \partial_s H_k(s) = \sigma_v^2 \mathbb{E}[\rho'(P_k(s)) \mathbb{E}[H_k(s) \nabla_{\mathbf{z}^b} F_k^{\mathbf{H},\mathbf{B}}(s, \mathbf{Z}^h, \mathbf{Z}^b) | \mathbf{Z}^h, \mathbf{Z}^b]] \cdot \mathbf{B}_{\wedge k-1}(s) \\ \qquad\qquad + \mathbb{E}[\rho(P_k(s)) G_k^{\mathbf{H},\mathbf{B}}(s, \mathbf{Z}^h, \mathbf{Z}^b) | \mathbf{H}, \mathbf{B}], \\ y_k := \mathbb{E}[H_k(1) W_{\text{out}}]. \end{cases}$$

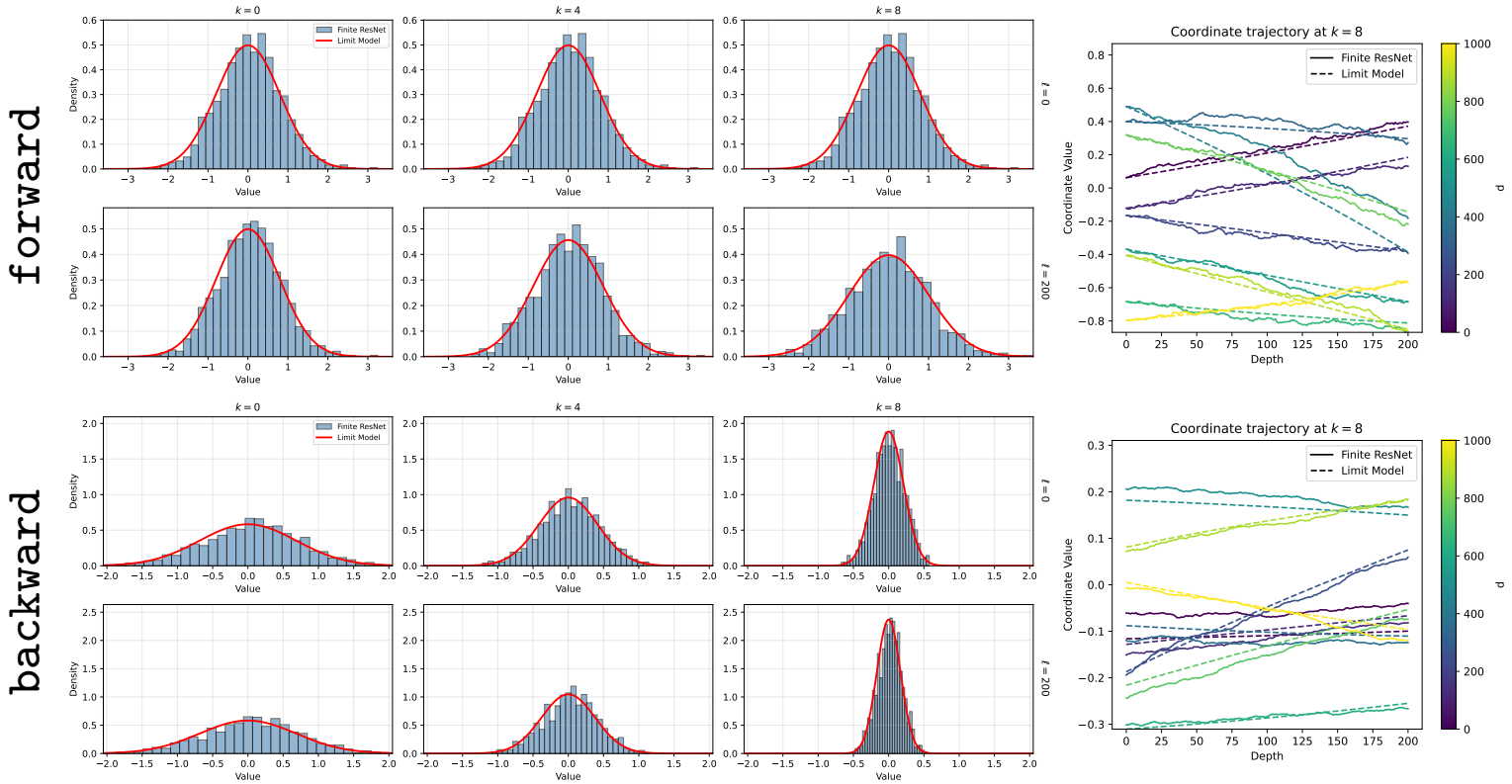
Backward pass:

$$\begin{cases} Q_k(s) = Z_k^b(s) + \mathbb{E}[B_k(s) G_k^{\mathbf{H},\mathbf{B}}(s, \mathbf{Z}^h, \mathbf{Z}^b) | \mathbf{Z}^h, \mathbf{Z}^b], \\ \partial_s B_k(s) = -\mathbb{E}\left[\rho'(P_k(s)) F_k^{\mathbf{H},\mathbf{B}}(s, \mathbf{Z}^h, \mathbf{Z}^b) Q_k(s) | \mathbf{H}, \mathbf{B}\right] \\ \qquad -\sigma_u^2 \mathbb{E}\left[\rho'\left(P_k(s)\right) \mathbb{E}[B_k(s) \nabla_{\mathbf{z}^h} G_k^{\mathbf{H},\mathbf{B}}(s, \mathbf{Z}^h, \mathbf{Z}^b) | \mathbf{Z}^h, \mathbf{Z}^b]\right] \cdot \mathbf{H}_{\wedge k-1}(s) \\ \qquad -\sigma_u^2 \mathbb{E}\Big\{\rho''(P_k(s)) Q_k(s) \Big[H_k(s) \\ \qquad +\mathbb{E}[H_k(s) \nabla_{\mathbf{z}^h} F_k^{\mathbf{H},\mathbf{B}}(s, \mathbf{Z}^h, \mathbf{Z}^b) | \mathbf{Z}^h, \mathbf{Z}^b] \cdot \mathbf{H}_{\wedge k-1}(s)\Big] | \mathbf{H}, \mathbf{B}\Big\}, \\ B_k(1) = W_{\text{out}} \cdot \nabla \ell(y_k), \end{cases}$$

Gaussian $(\mathbf{Z}^h, \mathbf{Z}^b) \in L^2([0,1], \mathbb{R}^{2K})$ with same covariance as $(U\mathbf{H}, V\mathbf{B})$.

Linear setting

In the linear case $\rho(z) = z$, the limit is explicit.



→ This matches the behaviour of finite ResNets.

DMFT Closure

We only need evaluation at $(\mathbf{Z}^h(s), \mathbf{Z}^b(s))$. Setting

$$\begin{cases} (F_k, G_k) := (F_k^{\mathbf{H},\mathbf{B}}(\mathbf{Z}^h, \mathbf{Z}^b), G_k^{\mathbf{H},\mathbf{B}}(\mathbf{Z}^h, \mathbf{Z}^b)), \\ (\nabla_{\mathbf{z}^\mathrm{j}} F_k, \nabla_{\mathbf{z}^\mathrm{j}} G_k) := (\nabla_{\mathbf{z}^\mathrm{j}} F_k^{\mathbf{H},\mathbf{B}}(\mathbf{Z}^h, \mathbf{Z}^b), \nabla_{\mathbf{z}^\mathrm{j}} G_k^{\mathbf{H},\mathbf{B}}(\mathbf{Z}^h, \mathbf{Z}^b)), \ \mathrm{j} \in \{h, b\}, \end{cases}$$

we have the update rules

$$\begin{cases} F_{k+1} = F_k - \eta_u \rho'(P_k) Q_k H_k, \\ G_{k+1} = G_k - \eta_v \rho(P_k) B_k, \end{cases}$$

$$\nabla_{\mathbf{z}^\mathrm{j}} F_{k+1} = \begin{pmatrix} \nabla_{\mathbf{z}^\mathrm{j}} F_k \\ 0 \end{pmatrix} - \eta_u \left(\rho''(P_k) Q_k \begin{pmatrix} \mathbb{E}[H_k \nabla_{\mathbf{z}^\mathrm{j}} F_k | \mathbf{z}^h, \mathbf{z}^b] \\ \mathbb{1}_{\mathrm{j}=h} \end{pmatrix} + \rho'(P_k) \begin{pmatrix} \mathbb{E}[B_k \nabla_{\mathbf{z}^\mathrm{j}} G_k | \mathbf{z}^h, \mathbf{z}^b] \\ \mathbb{1}_{\mathrm{j}=b} \end{pmatrix} \right) H_k,$$

$$\nabla_{\mathbf{z}^\mathrm{j}} G_{k+1} = \begin{pmatrix} \nabla_{\mathbf{z}^\mathrm{j}} G_k \\ 0 \end{pmatrix} - \eta_v \rho'(P_k) \begin{pmatrix} \mathbb{E}[H_k \nabla_{\mathbf{z}^\mathrm{j}} F_k | \mathbf{z}^h, \mathbf{z}^b] \\ \mathbb{1}_{\mathrm{j}=h} \end{pmatrix} B_k, \quad \text{for } \mathrm{j} \in \{h, b\}.$$

$\hookrightarrow$ Closure in the variables $(W_{\mathrm{in}}, W_{\mathrm{out}}, \mathbf{H}, \mathbf{B}, \mathbf{F}, \mathbf{G}, \nabla\mathbf{F}, \nabla\mathbf{G})$.

$\hookrightarrow$ Similar to linear response functions in DMFT [Bordelon et al. '22, Montanari et al.'25,...].

Mean-field limit

Theorem (C.,Chizat,Maass'26)

The limit system admits a unique solution $(\mathbf{H}, \mathbf{B}) \in L^2(\Omega, C([0,1], \mathbb{R}^{2K}))$.

Theorem (C.,Chizat,Maass'26)

Let h_k^D *and* H_k^D *be coupled through the same* $(W_{\text{in}}^D, W_{\text{out}}^D)$.

$\exists C_k > 0 : \forall D \geq 1, \forall \delta \in (e^{-D}, 1)$,

$$\sup_{s \in [0,1]} \left\| h_k^D(s) - H_k^D(s) \right\|_{\overline{D}} \leq C_k \frac{1 + \log(1/\delta)}{\sqrt{D}},$$

with proba $\geq 1 - \delta$.

If furthermore $(W_{\text{in}}^D, W_{\text{out}}^D)$ *is a.s. bounded, then this holds for* $\delta \in (0,1)$, *yielding* L^2*-convergence.*

Cavity method

Taylor expansion:

$$\mathbb{E}[\rho(\langle U_k^D, h_k^D\rangle_{\overline{D}})\sqrt{D}V^d|\mathbf{h},\mathbf{b}]$$
$$= \mathbb{E}[\rho(S_k^{\mathbf{h}^D} + \langle f_k^{\mathbf{h}^D,\mathbf{b}^D}(\mathbf{S}^{\mathbf{h}^D},\mathbf{S}^{\mathbf{b}^D}), h_k^D\rangle_{\overline{D}})\sqrt{D}V^d|\mathbf{h},\mathbf{b}]$$
$$\simeq \mathbb{E}[\rho(S_k^{\mathbf{h}^D} + \langle f_k^{\mathbf{h}^D,\mathbf{b}^D}(\mathbf{S}^{\mathbf{h}^D},\overline{\mathbf{S}}^{\mathbf{b}^D}), h_k^D\rangle_{\overline{D}})\sqrt{D}V^d|\mathbf{h},\mathbf{b}]$$
$$+ \mathbb{E}[\rho'(S_k^{\mathbf{h}^D} + \langle f_k^{\mathbf{h}^D,\mathbf{b}^D}(\mathbf{S}^{\mathbf{h}^D},\overline{\mathbf{S}}^{\mathbf{b}^D}), h_k^D\rangle_{\overline{D}})\langle \nabla_{\mathbf{z}^b} f_k^{\mathbf{h}^D,\mathbf{b}^D}(\mathbf{S}^{\mathbf{h}^D},\overline{\mathbf{S}}^{\mathbf{b}^D})[V^d\mathbf{b}^d], h_k^D\rangle_{\overline{D}} V^d]$$
$$+ O(D^{-1/2}),$$

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

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

As $D \to +\infty$, we get

$$\sigma_v^2\,\mathbb{E}[\rho'(P_k)\mathbb{E}[H_k \nabla_{\mathbf{z}^b} F_k^{\mathbf{H},\mathbf{B}}(\mathbf{Z}^{\mathbf{H}},\mathbf{Z}^{\mathbf{B}})|\mathbf{Z}^{\mathbf{H}},\mathbf{Z}^{\mathbf{B}}]|\mathbf{H},\mathbf{B}]\mathbf{B}.$$

+ Quantitative CLT.

Conclusion

Highlights: a theory of ResNets in the large-scale limit

- Phase diagram of ResNets.
- Infinite depth ($L \to +\infty$) implies infinite width ($M \to +\infty$).
- High-dimensional ($D \to +\infty$) limit of Mean ODE.
- Joint limit at rate $O\big(\frac{1}{L} + \sqrt{\frac{D}{LM}} + \frac{1}{\sqrt{D}}\big)$ – numerically sharp.

Perspectives: complete phase diagram, learning properties of the limit, tighten connections to practice...

Articles:

- Chizat (2025). The Hidden Width of Deep ResNets: Tight Error Bounds and Phase Diagrams <arXiv:2509.10167>.
- C., Chizat, Maass (2026). ResNets of All Shapes and Sizes: Convergence of Training Dynamics in the Large-scale Limit <arXiv:2603.18168>.