Topics in ML, Lecture 12
ResNets (II)

Lénaïc Chizat*

December 15, 2025

In today's class, we study the training dynamics of ResNets in the *local feature learning* regime – which is optimal in practice – and prove a quantitative convergence to a limit dynamics in the large depth limit. The content is based on [Chizat, 2025, Section 2].

Contents

1 Residual Neural Networks: reminders

We recall the definitions from last week, specialized to the *local feature learning* scaling—the one which is optimal in practice—with effective scale of the residual block $\alpha = \frac{1}{ML}$. Consider a ResNet with embedding dimension d, hidden width M and depth L defined as follows. For an input $x \in \mathbb{R}^d$,

$$\hat{h}_\theta^0 = x, \qquad \hat{h}_\theta^\ell = \hat{h}_\theta^{\ell-1} + \frac{1}{LM}\sum_{j=1}^{M} \phi(\hat{h}_\theta^{\ell-1}, z^{j,\ell}), \qquad \ell \in [1:L] \tag{1}$$

where $\theta = (z^{j,\ell})_{j,\ell} \in (\mathbb{R}^p)^{M\times L}$ are the parameters, $\phi : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^d$ is a nonlinear map.

We consider, to avoid lengthy expressions, a single sample $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ and the objective function

$$\hat{\mathcal{L}}(\theta) = \mathrm{loss}(y, \hat{h}_\theta^L).$$

*EPFL lenaic.chizat@epfl.ch

We recall from last week that $\forall j \in [1:M]$, $\forall \ell \in [1:L]$,

$$\nabla_{z^{j,\ell}} \hat{\mathcal{L}}_i(\theta) = \frac{1}{ML} D_2\phi\big(\hat{h}_\theta^{\ell-1},\, z^{j,\ell}\big)^\top \hat{b}_\theta^\ell\big(\nabla \mathrm{loss}(y, \hat{h}_\theta^L)\big), \tag{2}$$

where the backward pass $(\hat{b}_\theta^\ell)_{\ell\in[1:L]} \in (\mathbb{R}^d)^L$ is characterized by the backward recursion

$$\hat{b}_\theta^L(w) = w, \qquad \hat{b}_\theta^{\ell-1}(w) = \hat{b}_\theta^\ell(w) + \frac{1}{LM}\sum_{j=1}^M D_1\phi\big(\hat{h}_\theta^{\ell-1},\, z^{j,\ell}\big)^\top \hat{b}_\theta^\ell(w). \tag{3}$$

We consider an initial probability distribution $\mu_0 \in \mathcal{P}(\mathbb{R}^p)$ and a learning-rate ηLM (this is the suitable scaling in this context, cf lecture on two-layer NNs). The gradient descent (GD) dynamics $(\theta_k)_{k\geq 0} = (\hat{Z}_k^{j,\ell})_{j,\ell,k}$ is given by

$$\begin{aligned} \hat{Z}_0^{j,\ell} \overset{\text{iid}}{\sim} \mu_0, \quad \hat{Z}_{k+1}^{j,\ell} &= \hat{Z}_k^{j,\ell} - \eta ML \nabla_{z^{j,\ell}} \mathcal{L}(\theta_k), && \forall j \in [1:M],\ \forall \ell \in [1:L],\ \forall k \in \mathbb{N}. \\ &= \mathrm{Update}(\hat{Z}_k^{j,\ell}, \hat{h}_k^{\ell-1}, \hat{b}_k^\ell(\hat{g}_k)), && \hat{g}_k = \nabla \mathrm{loss}(y, \hat{h}_k^L) \end{aligned}$$

where the map $\mathrm{Update} : \mathbb{R}^p \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^p$ is defined as

$$\mathrm{Update}(z, h, b) = z - \eta D_2\phi(h, z)^\top b$$

We will make appropriate assumptions so that this map is Lipschitz continuous to simplify the argument.

2 Limit model: Neural Mean ODE

We now present the limit model, which we refer to as the (forward) Mean ODE.

2.1 Informal derivation

Let us first motivate the limit model from an intuitive viewpoint. Let $\bar{h}_k^\tau(s)$ for $s \in [0,1]$ be the piecewise linear interpolation of the forward pass at GD step k with step-size $\tau = 1/L$, ie such that $\bar{h}_k^\tau(s) = \hat{h}_k^\ell$ for $s \in [\tau\ell, \tau(\ell+1)[$, $\ell \in [0:L-1]$ and $h_k^\tau(1) = \hat{h}_k^L$. Then it holds

$$\frac{\bar{h}_k^\tau(s_\ell + \tau) - \bar{h}_k^\tau(s_\ell)}{\tau} = \frac{1}{M}\sum_{j=1}^M \phi(\bar{h}_k^\tau(s_\ell), z^{j,\ell+1})$$

As $L \to \infty$ (hence $\tau \to 0$), one may expect that the left-hand side converges to a derivative $\partial_s h_k(s)$. As $M \to \infty$, one may expect that the right-hand side converges to the expectation $\mathbf{E}[\phi(h(s), Z(s))]$ for some random variable $Z(s)$, at least if the $z^{j,\ell+1}$ are asymptotically independent (which is not obvious at first sight).

Both of these guesses, when suitably interpreted, turn out to be correct. There is in fact something more surprising happening: this limit arises as soon as $L \to \infty$, even when M does not diverge (e.g. $M = 1$), thanks to the fact that averaging also takes place "across depth", not just "across width". This effect is studied in Section 4.

2.2 Rigorous definition

We parameterize the limit model by a L^2 map $Z : [0,1] \times \Omega \to \mathbb{R}^p$ where $(\Omega, \mathbf{P})$ is an abstract probability space. We may interpret Z as a stochastic process indexed by a depth index $s \in [0,1]$ whose marginal distributions $\mathrm{Law}(Z(s))$ represents the distribution of parameters at

this layer. The forward pass $h_Z(s) \in \mathbb{R}^D$ is a function of depth $s \in [0,1]$, input $x \in \mathbb{R}^d$ (which is not in notation since we fix it throughout) and the stochastic process Z that encodes the parameters of the limit model. It is characterized as the solution to the forward Mean ODE:

$$h_Z(0) = x, \qquad \partial_s h_Z(s) = \mathbf{E}\big[\phi(h_Z(s), Z(s))\big], \quad \forall s \in [0,1], \forall x \in \mathbb{R}^D. \tag{4}$$

Note that h_Z only depends on Z through the distribution of the marginals $(\mathrm{Law}(Z(s)))_{s\in[0,1]}$. However, it is more convenient (in my opinion) to work with the stochastic process representation Z rather than with a family of measures $(\mu_s)_{s\in[0,1]}$.

Similarly as above, we consider a single sample (x, y) and the objective is defined as

$$\mathcal{L}(Z) = \mathrm{loss}(y, h_Z(1))$$

and we consider GD of $\mathcal{L}$ in the L^2 geometry starting from a random constant:

$$Z_0 \sim \mu_0, \qquad Z_{k+1} = Z_k - \eta \nabla \mathcal{L}(Z_k), \quad \forall k \in \mathbb{N}. \tag{5}$$

(here, with a slight abuse of notation, $Z_0 \sim \mu_0$ means that $Z_0(s) = Z_0$ is independent[1] of s and $\mathrm{Law}(Z_0(0)) = \mu_0$.) Observe that this is a deterministic dynamics in $L^2([0,1] \times \Omega; \mathbb{R}^p)$.

2.3 Expression of the gradient and dynamics

The gradient's expression can be derived from the adjoint method (i.e. continuous-time backpropagation). The backward mean ODE $b_Z(s, w) \in \mathbb{R}^d$ with $s \in [0,1]$ and $w \in \mathbb{R}^d$ is the solution to

$$b_Z(1, w) = w, \qquad \partial_s b_Z(s, w) = -\mathbf{E}\Big[D_1\phi(h_Z(s), Z(s))^\top b_Z(s, w)\Big], \ s \in [0,1]. \tag{6}$$

One has the following equations for the GD dynamics $(Z_k)_{k\geq 0}$, $\forall s \in [0,1], \forall k \geq 0$:

$$Z_0 \sim \mu_0, \tag{7}$$

$$Z_{k+1}(s) = Z_k(s) - \eta D_2\phi(h_k(s), Z_k(s))^\top b_k(s, \nabla \mathrm{loss}(h_k(1))) \tag{8}$$

$$= \mathrm{Update}(Z_k(s), h_k(s), b_k(s, g_k)), \qquad g_k = \nabla \mathrm{loss}(y, h_k(1)) \tag{9}$$

where the Update map is the same as above and $(h_k(s), b_k(s,w)) = (h_{Z_k}(s), b_{Z_k}(s,w))$. The rigorous connection between this dynamics and the ResNet dynamics is the object of the following sections.

3 Mean ODEs and some regularity results

In the limit model, the forward and backward passes are expressed as Mean ODEs, that is ODEs where the right-hand side is given by an expectation. Let us discuss some properties of these objects.

3.1 Well-posedness

Consider a generic *Mean ODE* of the form

$$a(0) \in \mathbb{R}^d, \qquad \dot{a}(s) = F(s, a(s)), \qquad F(s,x) := \mathbf{E}[f(s, x, Z(s))] \tag{10}$$

where $f : [0,1] \times \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^d$ and $(Z(s))_{s\in[0,1]}$ is a stochastic process.

[1] Our choice to initialize with a constant is a convenient convention: it will allow us to control the regularity in s of the ODE (4) associated to Z_k in terms of the regularity of $s \mapsto Z_k(s)$, which is easy to track.

Lemma 3.1. *Assume that there exists $L > 0$ such that f is L-Lipschitz and that $s \mapsto Z(s)$ is almost surely L-Lipschitz. Then for any $a(0) \in \mathbb{R}^d$, (10) admits a unique solution $a : [0,1] \to \mathbb{R}^d$ and there exists $R > 0$ only dependent on $\|a(0)\|_2$ and L such that $\|a(s)\|_2 \leq R$ and $s \mapsto a(s)$ is R-Lipschitz continuous.*

Proof. Under our assumptions, F is continuous in s (in fact Lipschitz), uniformly Lipschitz in x and has at most linear growth in x so the mean ODE has a unique global solution on $[0,1]$ by Picard-Lindelöf's theorem. Using Grönwall's lemma, we deduce bounds on $\|a(s)\|_2$ and $\|\dot{a}(s)\|_2$. □

3.2 Propagation of Lipschitz regularity and subgaussian tails

We now come back to the training dynamics of the limit model and derive controls on the regularity of the maps $s \mapsto h_k(s)$, $s \mapsto b_k(s)$ and $s \mapsto Z_k(s)$. For this, it will be useful to remark that

- the forward pass at iteration k is a Mean ODE of the form (10) with $a(0) = x$ and $f = f_k^h : (s, a, z) \mapsto \phi(a, z)$;
- the backward pass at iteration k, after "depth-reversal" (that is, composition with $s \mapsto 1-s$) is a Mean ODE of the form (10) with $a(0) = g_k$ and $f = f_k^b : (s, a, z) \mapsto D_1\phi(h_k(s), z)^\top a$.

Lemma 3.2 (Lipschitz regularity)**.** *Assume that ϕ and $D\phi$ are Lipschitz continuous. Then for $k \geq 0$, it holds*

(i) the sequences $(Z_k)_k$, $(h_k)_k$, $(g_k)_k$, $(b_k)_k$ are uniquely well defined by the GD equations;

(ii) the functions $h_k(\cdot)$ and $b_k(\cdot, g_k)$ (of type $[0,1] \to \mathbb{R}^d$) are Lipschitz.

(iii) there exists $c_k > 0$ such that the map $s \mapsto Z_k(s)$ is c_k-Lipschitz, almost surely.

Proof. By recursion (exercice). □

Lemma 3.3 (Subgaussian tails)**.** *Assume that ϕ and $D\phi$ are Lipschitz and that $Z_0 \sim \mu_0$ is σ_0^2-subgaussian. Then $\forall k \geq 0$, there exists $c_k > 0$ (independent of σ_0) such that $\forall s \in [0,1]$, $Z_k(s)$ and $f_k^h(s, h_k(s), Z_k(s))$ and $f_k^b(s, b_k(s, g_k), Z_k(s))$ are $(c_k\sigma_0^2)$-subgaussian.*

Proof. By recursion (exercice). □

4 Stochastic approximation of Mean ODEs

Proposition 4.1. *Let f and Z and a be as in Lemma 3.1 and such that $\forall s \in [0,1]$, $f(s, a(s), Z(s))$ is σ^2-subgaussian. For integers $M, L \geq 1$, let $s_\ell = \ell/L$ and integrate the mean ODE with the following "Euler Monte-Carlo" scheme :*

$$\hat{a}^0 \in \mathbb{R}^d, \qquad \hat{a}^\ell = \hat{a}^{\ell-1} + \frac{1}{LM}\sum_{j=1}^{M} f(s_{\ell-1}, \hat{a}^{\ell-1}, \hat{Z}^{j,\ell}), \qquad \ell \in [1:L] \tag{11}$$

where $(\hat{Z}^{j,\ell})_{j,\ell}$ are random vectors such that there exists a family of $M \times L$ independent copies $Z^{j,\ell}$ of Z such that $\|\hat{Z}^{j,\ell} - Z^{j,\ell}(s_{\ell-1})\|_2 \leq \epsilon$ almost surely. Then there exists $c > 0$ such that $\forall \delta \in]0,1]$, with probability at least $1-\delta$, it holds

$$\sup_{\ell \in [1:L]} \|\hat{a}^\ell - a(s_\ell)\|_2 \leq c\Big(\epsilon + \|a(0) - \hat{a}^0\|_2 + \frac{1}{L} + \frac{\sigma(1 + \sqrt{\log(1/\delta)})}{\sqrt{ML}}\Big). \tag{12}$$

Remark 4.2 (Extension to inexact model)**.** *One can also consider an inexact model in Eq. (11) that uses a function $\hat{f}$ which is such that $\|f - \hat{f}\|_\infty \leq \epsilon'$. Then a straightforward extension of the proof leads to the same conclusion with an extra ϵ' term inside the parentheses of Eq. (12).*

Proof. For $\ell \in [0 : L-1]$, it holds

$$\begin{aligned}
a(s_{\ell+1}) - \hat{a}^{\ell+1} &= a(s_\ell) - \hat{a}^\ell + \int_{s_\ell}^{s_{\ell+1}} \dot{a}(s)\mathrm{d}s - \frac{1}{ML}\sum_{j=1}^{M} f(s_\ell, \hat{a}^\ell, \hat{Z}^{j,\ell+1}) \\
&= a(s_\ell) - \hat{a}^\ell + \underbrace{\int_{s_\ell}^{s_{\ell+1}} \dot{a}(s)\mathrm{d}s - \frac{1}{L}F(s_\ell, a(s_\ell))}_{e_{euler}^{\ell+1}} \\
&\quad + \underbrace{\Big(\frac{1}{L}F(s_\ell, a(s_\ell)) - \frac{1}{ML}\sum_{j=1}^{M} f(s_{\ell-1}, a(s_\ell)), Z^{j,\ell+1}(s_\ell))\Big)}_{e_{mc}^{\ell+1}} \\
&\quad + \underbrace{\frac{1}{ML}\sum_{j=1}^{M}\Big(f(s_\ell, a(s_\ell), Z^{j,\ell+1}(s_\ell)) - f(s_\ell, \hat{a}^\ell, \hat{Z}^{j,\ell+1})\Big)}_{e_{approx}^{\ell+1}}
\end{aligned}$$

By recursion, we have

$$a(s_\ell) - \hat{a}^\ell = a(0) - \hat{a}^0 + \sum_{k=1}^{\ell} e_{euler}^k + \sum_{k=1}^{\ell} e_{mc}^k + \sum_{k=1}^{\ell} e_{approx}^k$$

and therefore, with $\Delta_\ell \coloneqq \|a(s_\ell) - \hat{a}^\ell\|_2$, it holds

$$\Delta_\ell \leq \|a(0) - \hat{a}^0\|_2 + \sum_{k=1}^{\ell} \|e_{euler}^k\|_2 + \Big\|\sum_{k=1}^{\ell} e_{mc}^k\Big\|_2 + \sum_{k=1}^{\ell} \|e_{approx}^k\|_2. \tag{13}$$

Note that for the Monte-Carlo error term, we take the norm *after* summing across layers. Let us bound these error terms one by one. First, using the Lipschitz continuity of f, a and Z,

$$\begin{aligned}
\|e_{euler}^{\ell+1}\|_2 &= \Big\|\int_{s_\ell}^{s_{\ell+1}} \Big(F(s, a(s)) - F(s_\ell, a(s_\ell))\Big)\mathrm{d}s\Big\|_2 \\
&\leq \int_{s_\ell}^{s_{\ell+1}} \mathbf{E}\Big[\|f(s, a(s), Z(s)) - f(s_\ell, a(s_\ell), Z(s_\ell))\|_2\Big]\mathrm{d}s \\
&\lesssim \int_{s_\ell}^{s_{\ell+1}} |s - s_\ell|\mathrm{d}s \lesssim \frac{1}{L^2}.
\end{aligned}$$

Moreover, using the Lipschitz continuity of f,

$$\|e_{approx}^{\ell+1}\|_2 \lesssim \frac{\Delta_\ell + \epsilon}{L}.$$

Finally, the random vectors $(e_{mc}^\ell)_{\ell=1}^L$ are independent, centered and subgaussian, with variance proxy $\sigma^2/(L^2M)$. It follows that $\sum_{k=1}^{\ell} e_{mc}^k$ is centered and subgaussian with variance proxy $\sigma^2\ell/(L^2M)$. By multidimensional subgaussian concentration (see exercice sheet), it holds with probability at least $1-\delta$,

$$\Big\|\sum_{k=1}^{\ell} e_{mc}^\ell\Big\|_2 \lesssim \sqrt{\frac{\sigma^2\ell}{L^2M}}(1 + \sqrt{\log(1/\delta)}) \leq \frac{\sigma}{\sqrt{ML}}(1 + \sqrt{\log(1/\delta)}).$$

By Lévy-Ottaviani inequality (see Lemma A.1 in appendix), a similar bound holds for $\max_{\ell \le L} \left\| \sum_{k=1}^{\ell} e_{mc}^{\ell} \right\|_2$ (alternatively, one may use a union bound over $\ell \in [1:L]$ to control the max, but that creates a $\sqrt{\log(L)}$ factor which is not needed).

Plugging all these bounds into Eq. (13), we obtain that with probability at least $1-\delta$ for $\ell \in [1:L]$, it holds

$$\Delta_\ell \lesssim \|a(0) - \hat{a}^0\|_2 + \epsilon + \frac{\sigma(1+\sqrt{\log(1/\delta)})}{\sqrt{ML}} + \frac{1}{L} + \frac{1}{L}\sum_{k=0}^{\ell-1} \Delta_k.$$

The result follows by the discrete Grönwall lemma (integral form, see exercice sheet). □

5 Main convergence result

We are now ready to state the main result, that gives the quantitative convergence of the ResNet's training dynamics to the limit dynamics in the infinite depth limit.

Theorem 5.1. *Suppose that μ_0 is σ_0^2-subgaussian and that ϕ, $D\phi$ and ∇loss are Lipschitz. Let $(Z_k^{j,\ell})_{k\ge 0}$ be iid samples from the limit dynamics such that $Z_0^{j,\ell}(s) = \hat{Z}_0^{j,\ell}$ $\forall s \in [0,1]$ and consider*

$$\begin{aligned}
\Delta_k^Z &:= \sup_{j,\ell} \|Z_k^{j,\ell}(s_{\ell-1}) - \hat{Z}_k^{j,\ell}\|_2, \\
\Delta_k^h &:= \sup_{\ell} \|h_k(s_\ell) - \hat{h}_k^\ell\|_2, \\
\Delta_k^b &:= \sup_{\ell} \|b_k(s_\ell, g_k) - \hat{b}_k^\ell(\hat{g}_k)\|_2.
\end{aligned}$$

Then for $k \in \mathbb{N}^*$, there exists $c_k > 0$ (depending on k, d, and the characteristics of ϕ) such that with probability at least $1-\delta$ it holds

$$\max\{\Delta_k^Z, \Delta_k^h, \Delta_k^b\} \le c_k\Big(\frac{1}{L} + \frac{\sigma_0(1+\sqrt{\log(1/\delta)})}{\sqrt{ML}}\Big).$$

We can make the following remarks:

- This theorem proves convergence, for bounded training time, of the training dynamics of ResNets to the training dynamics of the limit model "Neural Mean ODE". A sufficient condition for convergence is $L \to \infty$ (irrespective of the scaling of M);
- One can verify empirically that the dependency in M and L is tight.
- If one tracks the dependency in d, say in the 2LP case, one finds that the critical residual scale is $\alpha = \Theta(\frac{\sqrt{d}}{ML})$ and in this regime the bound is in $O(\frac{1}{L} + \frac{\sqrt{d}}{\sqrt{ML}})$. In practice, d and M are comparable, so as L grows we indeed have convergence to the limit model.

Proof. • **Step 1: set-up.** Recall from Section 1, that it holds

$$\hat{Z}_{k+1}^{j,\ell} = \text{Update}(\hat{Z}_k^{j,\ell}, \hat{h}_k^\ell, \hat{b}_k^\ell(\hat{g}_k)), \qquad \hat{g}_k = \nabla\text{loss}(y, \hat{h}_k^L) \tag{14}$$

$$Z_{k+1}(s) = \text{Update}(Z_k(s), h_k(s), b_k(s, g_k)), \qquad g_k = \nabla\text{loss}(y, h_k(1)) \tag{15}$$

for the same Lipschitz map Update. Therefore, $\exists L > 0$ such that $\forall k \in \mathbb{N}^*$,

$$\Delta_{k+1}^Z \le L(\Delta_k^Z + \Delta_k^h + \Delta_k^b).$$

Let us now fix $k \in \mathbb{N}^*$ and control these various terms with high probability.

- **Step 2: Control on $\Delta_k(h)$.** Let us verify that we can use Proposition 4.1 with $f = f_k^h : (s,a,z) \mapsto \phi(a,z)$ and $Z = Z_k$. Clearly, f_k^h is Lipschitz and by Lemma 3.2, Z_k also. By Lemma 3.3, $f_k^h(s,a,Z_k(s))$ is subgaussian in s with variance proxy in $O(\sigma_0^2)$. Therefore the proposition applies (with $\epsilon = \Delta_k^Z$ and $\|a(0) - \hat{a}^0\|_2 = 0$) and there exists $c_{1,k}$ such that with probability at least $1-\delta$, it holds

$$\Delta_k^h \le c_{1,k}\Big(\Delta_k^Z + \frac{1}{L} + \sigma_0 \frac{1+\sqrt{\log(1/\delta)}}{\sqrt{ML}}\Big).$$

- **Step 3: Control on $\Delta_k(b)$.** Let us verify that we can use Proposition 4.1 with $f = f_k^b : (s,a,z) \mapsto D_1\phi(h_k(s),z)^\top a$ and $Z = Z_k$. By Lemma 3.2, one can deduce that f_k^b is Lipschitz and Z_k also. By Lemma 3.3, $f_k^b(s,a,Z_k(s))$ is uniformly subgaussian in s with variance proxy in $O(\sigma_0^2)$. Note also that we need the extension of Proposition 4.1 mentioned in Remark 4.2, with an inexact model $\hat{f} = \hat{f}_k^b : (s,a,z) \mapsto D_1\phi(\hat{h}_k(s),z)^\top a$. This modified proposition applies with $\epsilon = \Delta_k^Z$ and we get

$$\begin{aligned}\Delta_k^b &\le c\Big(\epsilon + \|f_k^b - \hat{f}_k^b\|_\infty + \|g_k - \hat{g}^k\|_2 + \frac{1}{L} + \sigma_0\frac{1+\sqrt{\log(1/\delta)}}{\sqrt{ML}}\Big)\\ &\le c\Big(\Delta_k^Z + \Delta_k^h + \frac{1}{L} + \sigma_0\frac{1+\sqrt{\log(1/\delta)}}{\sqrt{ML}}\Big)\end{aligned}$$

  where we have used $\|f_k^b - \hat{f}_k^b\|_\infty \lesssim \Delta_k^h$ and, since ∇loss is assumed Lipschitz, $\|g_k - \hat{g}^k\|_2 \lesssim \Delta_k^h$.

- **Step 4: Putting it all together.** Take a union bound over the $2k$ events such that all these bounds holds for all iterations before k (if we had n samples, then we would also take a union bound over the n samples, amounting to $2kn$ events). Then for all $0 \le k' \le k-1$, it holds, with probability at least $1-\delta$

$$\Delta_{k'+1}^Z \lesssim c\Big(\Delta_{k'}^Z + \frac{1}{L} + \frac{1+\sqrt{\log(2k/\delta)}}{\sqrt{ML}}\Big).$$

  Since $\Delta_0^Z = 0$, the conclusion for (Δ_k^Z) follows by Grönwall's lemma, and the conclusion for Δ_0^h and Δ_0^b by the controls derived in Step 2 and Step 3 respectively. □

References

Lénaïc Chizat. The hidden width of deep ResNets: Tight error bounds and phase diagrams. *arXiv preprint arXiv:2509.10167*, 2025.

Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

A Useful Lemma

Lemma A.1 (Lévy–Ottaviani inequality [De la Pena and Giné, 2012, Proposition 1.1.2])**.** *Let $X_1, \dots, X_L \in \mathbb{R}$ be independent random variables (not necessarily centered). Then for all $t > 0$,*

$$\mathbb{P}\left(\max_{1\le k\le L}\left\|\sum_{i=1}^k X_i\right\| > t\right) \le 3\max_{1\le k\le L}\mathbb{P}\left(\left\|\sum_{i=1}^k X_i\right\| > t/3\right).$$