A. Hyperparameters

We provide in Table 2 the training hyperparameters used in our numerical experiments. In Table 3, we give a short description of each hyperparameter. For the convolutional architecture, we also use a momentum of 0.9, a weight decay of 0.0005 and a cosine annealing learning rate scheduler (Loshchilov & Hutter, 2017).

Table 2: Training hyperparameters.

Dataset	Layers	N	B	η	$L_{\min}$	$L_{\max}$	$T_{\max}$	N_{epochs}	ϵ
Synthetic	Fully-connected	1,024	32	0.01	3	10,321	160	5	0.01
MNIST	Fully-connected	60,000	50	0.01	3	942	12,000	10	0.01
CIFAR-10	Convolutional	60,000	128	0.1	8	121	93,800	200	None

Table 3: Description of the values in Table 2. Note that $T_{\max} = \left\lceil \frac{N}{B} \right\rceil N_{\text{epochs}}$.

Parameter	Description
N	number of training samples
B	minibatch size
η	learning rate
$L_{\min}$	smallest network depth
$L_{\max}$	largest network depth
$T_{\max}$	max number of SGD updates
N_{epochs}	max number of epochs
ϵ	early stopping value

We report in Table 4 below results for $\tanh$ and trainable δ on the synthetic data with different batch sizes and learning rates (5 different seeds). We observe that the learning rate does affect α and β, but their sum always stays around 1. The batch size has no effect on the exponents.

Table 4: Average value of α (left) and β (right) for the trained weights, over 5 random initializations, η is the learning rate, and B the batch size.

α	$B = 8$	$B = 32$	$B = 128$
$\eta = .01$	$.69 \pm .02$	$.73 \pm .02$	$.67 \pm .02$
$\eta = .003$	$.59 \pm .05$	$.60 \pm .01$	$.58 \pm .01$
$\eta = .001$	$.58 \pm .01$	$.55 \pm .01$	$.53 \pm .01$

β	$B = 8$	$B = 32$	$B = 128$
$\eta = .01$	$.24 \pm .02$	$.29 \pm .05$	$.22 \pm .02$
$\eta = .003$	$.33 \pm .01$	$.41 \pm .06$	$.40 \pm .02$
$\eta = .001$	$.39 \pm .02$	$.43 \pm .02$	$.41 \pm .01$

B. Scaling Analysis of the Biases $b^{(L)}$ in the Fully-Connected Case

We mention in the main text that the behaviour of the trained values of $b^{(L)}$ with the depth L is similar to that of $A^{(L)}$, in the fully-connected case of Section 3.1. We verify these claims here. To do so, we follow the same methodology outlined in Section 2.3 for $b^{(L)}$, and we show the results for the $\tanh$ case in Figure 9 and for the ReLU case in Figure 10. We observe that the maximum norm, the scaled norm of the increments and the root sum of squares of $b^{(L)}$ scales in the same way as $A^{(L)}$ as the depth L increases. In particular, the scaling exponent β for $b^{(L)}$ is equal to the scaling exponent of $A^{(L)}$, justifying the setup considered in Section 4.1.

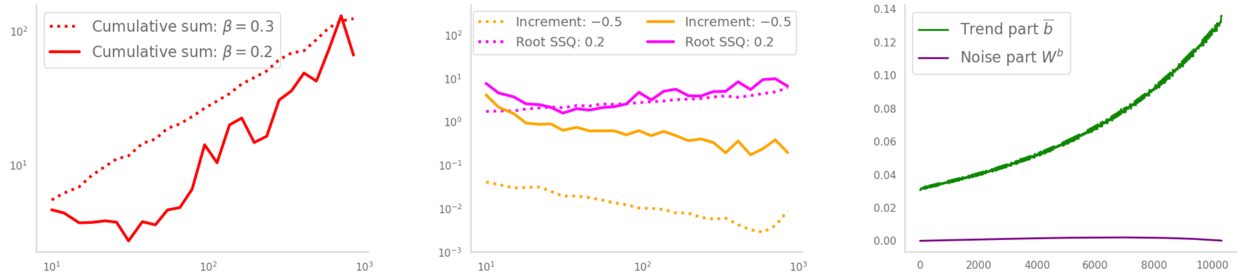

Figure 9: Scaling and hypothesis verification for $\tanh$ activation and $\delta^{(L)} \in \mathbb{R}$. Left: Maximum norm of $b^{(L)}$ with respect to L, in log-log scale. Middle: we plot in log-log scale the root sum of squares of $b^{(L)}$ in pink and the $\beta-$scaled norm of increments of $b^{(L)}$ in orange. The dashed lines are for the synthetic data and the solid lines are for MNIST. Right: Decomposition of the trained weights $b_{k,5}^{(L)}$ with the trend part $\bar{b}$ and the noise part W^b for $L = 10321$, as defined in (6), for the synthetic dataset.

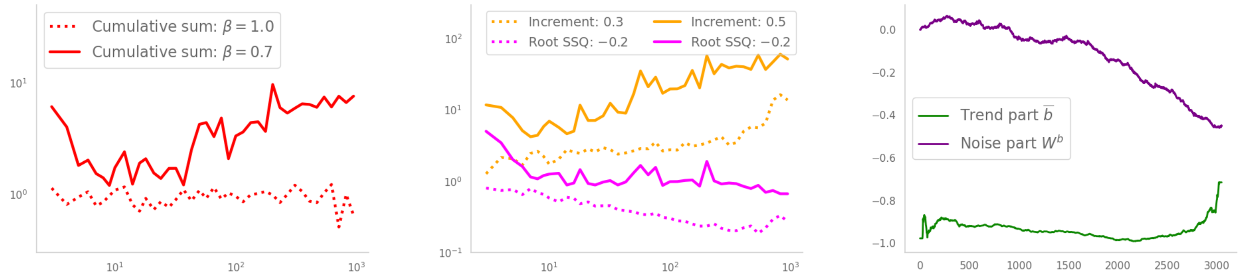

Figure 10: Scaling and hypothesis verification for ReLU activation and $\delta_k^{(L)} \in \mathbb{R}$. Left: Maximum norm of $|\delta^{(L)}|b^{(L)}$ with respect to L, in log-log scale. Middle: we plot in log-log scale the root sum of squares of $|\delta^{(L)}|b^{(L)}$ in pink and the $\beta-$scaled norm of increments of $|\delta^{(L)}|b^{(L)}$ in orange. The dashed lines are for the synthetic data and the solid lines for MNIST. Right: Decomposition of the trained weights $|\delta^{(L)}|\, b_{k,6}^{(L)}$ with the trend part $\bar{b}$ and the noise part W^b for $L = 10321$, as defined in (6), for the synthetic dataset.

C. Convolutional Network Results on CIFAR-10

We give in Table 5 the final test accuracy of our convolutional residual networks trained on an NVIDIA GeForce RTX 2080 GPU. The results are in line with those of traditional ResNet architectures (He et al., 2016), even though our networks do not have batch normalization layers (Ioffe & Szegedy, 2015). It is also noteworthy to add that our concept of depth is not that of traditional ResNets. We define the number of layers L as the number of skip connections in the network, that is the number of Δ_k kernels in (9).

We also note that the test error in Table 5 does not decrease with network depth. This is due to the fact, already mentioned in Section 3.3, that smaller depths usually suffice to get a good accuracy. In our case, we focus on a simple setting that still approaches the results obtained in practice. Rather than trying to find a setup that maximizes the accuracy, for instance with batch normalization or the Adam optimizer (Kingma & Ba, 2014), we aim to understand the scaling of residual networks in practical cases.

Table 5: Test error in $\%$ on CIFAR-10 for each network depth L.

L	8	11	12	14	16	20	24	28
Test error	6.64	6.37	6.32	5.98	6.25	5.98	6.24	7.03
L	33	42	50	65	80	100	121	
Test error	6.13	6.21	6.32	6.19	6.30	6.20	6.37	

D. Residual Network Architecture

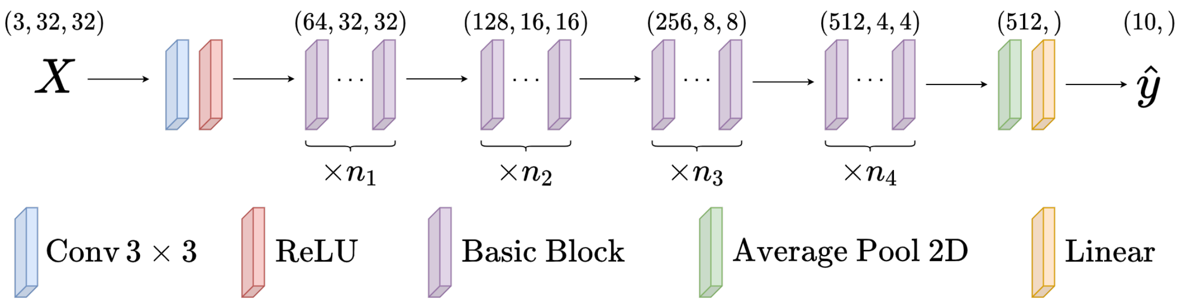

Figure 11: Residual architecture. There are 4 blocks that are respectively repeated n_1, n_2, n_3 and n_4 times. The network depth is $L = n_1 + n_2 + n_3 + n_4$. The Basic Block architecture is detailed in Figure 12.

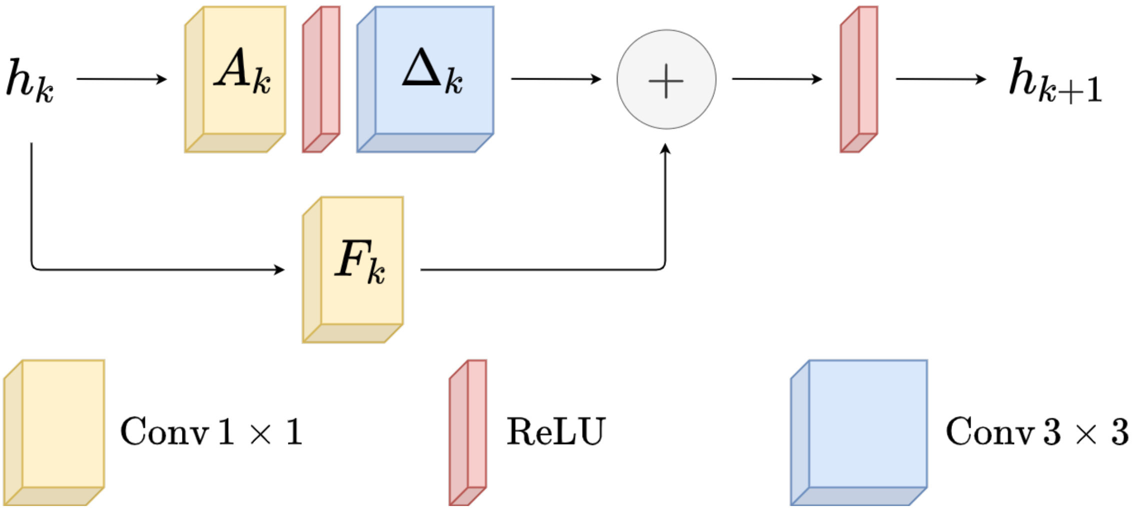

Figure 12: Basic Block from Figure 11. See (9) for details.

E. Proofs of Technical Results in Section 4

This appendix outlines the main arguments in the proofs of results stated in Section 4. Further mathematical details are provided in (Cohen et al., 2021).

E.1. Setup

As specified in Section 4.1, we model the cumulative sum of weights (resp. bias) as Itô processes $(W_t^A)_{t\geq 0}$ (resp. $(W_t^b)_{t\geq 0}$) on some filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$. This means W^A, W^b satisfy

$$\begin{aligned}\left(\mathrm{d}W_t^A\right)_{ij} &= \left(U_t^A\right)_{ij}\mathrm{d}t + \sum_{k,l=1}^{d}\left(q_t^A\right)_{ijkl}\left(\mathrm{d}B_t^A\right)_{kl} \quad \text{for } i,j = 1,\dots,d,\\ \mathrm{d}W_t^b &= U_t^b\mathrm{d}t + q_t^b\mathrm{d}B_t^b,\end{aligned} \tag{17}$$

where $(B_t^A)_{t\geq 0}$, resp. $(B_t^b)_{t\geq 0}$ are $d \times d$-dimensional, resp. d-dimensional, Brownian motions, and $q_t^A \in \mathbb{R}^{d,\otimes 4}$ and $q_t^b \in \mathbb{R}^{d\times d}$ for $t \in [0,1]$. We set $W_0^A = 0$, $W_0^b = 0$. Denote the quadratic variation processes as:

$$\begin{aligned}\left(\Sigma_t^A\right)_{i_1 j_1 i_2 j_2} &:= \sum_{k,l=1}^{d}\left(q_t^A\right)_{i_1 j_1 kl}\left(q_t^A\right)_{i_2 j_2 kl}, \quad \text{for } i_1, j_1, i_2, j_2 = 1,\dots,d,\\ \Sigma_t^b &:= q_t^b\left(q_t^b\right)^{\top}.\end{aligned} \tag{18}$$

We assume $(U_t^A)_{t\geq 0}$, $(U_t^b)_{t\geq 0}$, $(\Sigma_t^A)_{t\geq 0}$ and $(\Sigma_t^b)_{t\geq 0}$ are progressively measurable processes satisfying:

Assumption E.1 (Regularity assumptions)**.** *We assume*

(i) There exists a constant $C_1 > 0$ such that almost surely

$$\sup_{0\leq t\leq 1} \left\|U_t^A\right\| + \sup_{0\leq t\leq 1} \left\|U_t^b\right\| + \sup_{0\leq t\leq 1} \left\|\Sigma_t^A\right\| + \sup_{0\leq t\leq 1} \left\|\Sigma_t^b\right\| \leq C_1. \tag{19}$$

(ii) There exist $M > 0$ and $\kappa > 0$ such that $\forall s, t \in [0, 1]$ almost surely

$$\left\|U_t^A - U_s^A\right\|^2 + \left\|U_t^b - U_s^b\right\|^2 + \left\|\Sigma_t^A - \Sigma_s^A\right\|^2 + \left\|\Sigma_t^b - \Sigma_s^b\right\|^2 \leq M|t-s|^\kappa \tag{20}$$

$$\left\|\bar{A}_t - \bar{A}_s\right\|^2 + \left\|\bar{b}_t - \bar{b}_s\right\|^2 \leq M|t-s|^\kappa. \tag{21}$$

Lemma E.2 (Uniform integrability)**.** *Under Assumption E.1-(i), we have, for any $p_0 > 1$*

$$\mathbb{E}\left[\sup_{0\leq s\leq 1} \left\|W_s^A\right\|^{p_0}\right],\ \mathbb{E}\left[\sup_{0\leq s\leq 1} \left\|W_s^b\right\|^{p_0}\right] < \infty. \tag{22}$$

Lemma E.2 is proven by first applying Minkowski inequality to $\mathbb{E}\left[\sup_{0\leq s\leq 1}\left\|W_s^A\right\|^{p_0}\right]$ and then applying Burkholder-Davis-Gundy inequality to $\mathbb{E}\left[\sup_{0\leq s\leq 1}\left\|\left(\int_0^s \sum_{k,l=1}^d \left(q_t^A\right)_{ijkl}\left(\mathrm{d}B_t^A\right)_{kl}\right)_{i,j}\right\|^{p_0}\right]$.

Assumption E.3 (Uniform integrability)**.** *There exist $p_1 > 4$ and a constant C_0 such that for all L,*

$$\mathbb{E}\left[\sup_{0\leq k\leq L}\left\|h_k^{(L)}\right\|^{p_1}\right] \leq C_0. \tag{23}$$

E.2. Proof of Theorem 4.3

We now provide a sketch of the proof for Theorem 4.3 under Assumption 4.2, Assumption E.1 with $W^A \equiv 0$ and $W^B \equiv 0$, and Assumption E.3. The detailed proof can be found in a companion paper (Cohen et al., 2021). Under Assumption E.3, there exists $C_\infty > 0$ such that $\sup_{L\in\mathbb{N}}\max_{k=1,2,\ldots,L}\left\|h_k^{(L)}\right\| \leq C_\infty$. Denoting $\Delta h_k^L := h_{k+1}^L - h_k^L$ and $M_k^{(L)}(h) := \overline{A}_{t_k} h + \bar{b}_{t_k}$, from (13) we have

$$\Delta h_k^{(L)} := h_{k+1}^{(L)} - h_k^{(L)} = L^{-\alpha}\sigma\left(L^{-\beta}M_k^{(L)}(h_k^{(L)})\right).$$

For any vector $x \in \mathbb{R}^d$, denote $(x)_i$ as the i-th component of x. Further denote $\Delta h_k^{(L),i}$ and $M_k^{(L),i}$ the i-th element of $\Delta h_k^{(L)}$ and $M_k^{(L)}$, respectively. Applying a third-order Taylor expansion of σ around 0 using Assumption 4.2 we get

$$\Delta h_k^{L,i} = L^{-1}M_k^{(L),i}(h_k^{(L)}) + \frac{1}{2}\sigma''(0)L^{-\beta-1}\left(M_k^{(L),i}(h_k^{(L)})\right)^2 + \frac{1}{6}\sigma'''(\nu_k^i)L^{-2\beta-1}\left(M_k^{(L),i}(h_k^{(L)})\right)^3 \tag{24}$$

with $|\nu_k^i| \leq L^{-\beta}\left|\left(\overline{A}_{t_k}h_k^{(L)} + \bar{b}_{t_k}\right)_i\right|$ under the condition $\alpha + \beta = 1$. Denote $\{t_k = k/L,\ k = 0, 1, \ldots, L\}$ as the uniform partition of the interval $[0, 1]$. For $t \in [t_k, t_{k+1}]$, define

$$\widetilde{H}_t^{(L)} = h_k^{(L)} + (t - t_k)M_k^{(L),i}(h_k^{(L)}) + \frac{1}{2}\sigma''(0)L^{-\beta-1}\left(M_k^{(L),i}(h_k^{(L)})\right)^2 + \frac{1}{6}\sigma'''(\nu_k^i)L^{-2\beta-1}\left(M_k^{(L),i}(h_k^{(L)})\right)^3.$$

Then we have $\widetilde{H}_{t_k}^{(L)} = h_k^{(L)}$ for all $k = 0, 1, \cdots, L$.

Recall H_t the solution to the ODE (14). Denote the differences $d_k^{(L),1}(t) = \widetilde{H}_t^{(L)} - h_k^{(L)}$ and $d_k^{(L),2}(t) = H_t - \widetilde{H}_t^{(L)}$ for $t \in [t_k, t_{k+1}]$. Similarly denote the errors $e_k^{(L),1} = \sup_{t_k\leq t\leq t_{k+1}}\left\|d_k^{(L),1}(t)\right\|$ and $e_k^{(L),2} = \sup_{t_k\leq t\leq t_{k+1}}\left\|d_k^{(L),2}(t)\right\|$. The proof reduces to showing $\sup_{1\leq k\leq L} e_k^{(L),1} \to 0$ and $\sup_{1\leq k\leq L} e_k^{(L),2} \to 0$ when $L \to \infty$.

We first bound $e_k^{(L),1}$. Denote $c_0 := \sup_{x\in\mathbb{R}}|\sigma'''(x)| < \infty$. By definition and direct calculation, we have $e_k^{(L),1} \leq D_\infty L^{-1}$ with constant $D_\infty := A_{\max}C_\infty + b_{\max} + \frac{1}{2}\sigma''(0)(A_{\max}C_\infty + b_{\max})^2 + \frac{1}{6}c_0(A_{\max}C_\infty + b_{\max})^2$. Therefore it holds that $\lim_{L\to\infty}\sup_{1\leq k\leq L} e_k^{(L),1} = 0$.

We next bound $e_k^{(L),2}$. For $t \in [t_k, t_{k+1}]$,

$$\begin{aligned} d_{k+1}^{(L),2}(t) &= d_k^{(L),2}(t_{k+1}) - (t - t_{k+1}) M_{k+1}^{(L),i}(h_{k+1}^{(L)}) + \int_{t_{k+1}}^{t} (\overline{A}_s H_s + \bar{b}_s) \mathrm{d}s \\ &\quad -\frac{1}{2}\sigma''(0) L^{-\beta-1} \left(M_{k+1}^{(L),i}(h_{k+1}^{(L)}) \right)^2 - \frac{1}{6}\sigma'''(\nu_{k+1}^i) L^{-2\beta-1} \left(M_{k+1}^{(L),i}(h_{k+1}^{(L)}) \right)^3 \end{aligned} \tag{25}$$

From (24) and (25) we have

$$\begin{aligned} e_{k+1}^{(L),2} &\leq e_k^{(L),2} + \sup_{t_{k+1} \leq t \leq t_{k+2}} \left\| \int_{t_{k+1}}^{t} \left((\overline{A}_s H_s + \bar{b}_s) - \left(\overline{A}_{t_{k+1}} h_{k+1}^{(L)} + \bar{b}_{t_{k+1}} \right) \right) \mathrm{d}s \right\| \\ &\quad + \frac{1}{2} \left| \sigma''(0) \right| L^{-\beta-1} \left\| M_{k+1}^{(L)}(h_{k+1}^{(L)}) \right\|^2 + \frac{1}{6} c_0 L^{-2\beta-1} \left\| M_{k+1}^{(L)}(h_{k+1}^{(L)}) \right\|^3 . \end{aligned}$$

As $\beta > 0$, the last two terms are $o(L^{-1})$. Also, direct calculation yields, for L big enough and $A_{\max} := \sup_{0 \leq t \leq 1} \left\| \overline{A}_t \right\| < \infty$,

$$\sup_{t_{k+1} \leq t \leq t_{k+2}} \left\| \int_{t_{k+1}}^{t} \left((\overline{A}_s H_s + \bar{b}_s) - \left(\overline{A}_{t_{k+1}} h_{k+1}^{(L)} + \bar{b}_{t_{k+1}} \right) \right) \mathrm{d}s \right\| \leq 2 A_{\max} L^{-1} e_{k+1}^{(L),2} + o(L^{-1}).$$

Finally, $e_0 = \mathcal{O}(L^{-1})$, so by Grönwall's lemma, we also have $\sup_{1 \leq k \leq L} e_k^{(L),2} = \mathcal{O}(L^{-1})$.

□

E.3. Proof of Theorem 4.4

We provide a sketch of the proof for Theorem 4.4 under Assumptions (4.2), (E.1) and (E.3) for the case $\alpha = 0$ and $\beta = 1$. Other cases follow similarly. The detailed proof can be found in a companion paper (Cohen et al., 2021).

When $\alpha = 0$ and $\beta = 1$, we define the targeted SDE limit for the discrete scheme (10) as follows:

$$\mathrm{d}H_t = \mu(t, H_t)\mathrm{d}t + \mathrm{d}V_t^A \, H_t + \mathrm{d}V_t^b, \quad 0 \leq t \leq 1, \ \text{ with } \ H_0 = x, \tag{26}$$

in which

$$\mu\left(t, h\right) = U_t^A \, h + \, U_t^b + \overline{A}_t h + \bar{b}_t + \frac{1}{2}\sigma''(0) Q(t,h), \quad \mathrm{d}V_t^A = \sum_{k,l=1}^{d} \left(q_t^A\right)_{ijkl} \left(\mathrm{d}B_t^A\right)_{kl}, \quad \mathrm{d}V_t^b = q_t^b \, \mathrm{d}B_t^b, \tag{27}$$

with $V_0^A = 0$ and $V_0^b = 0$. Here the quadratic variation process $\frac{1}{2}\sigma''(0)Q(t,h)$ is the *Itô correction* term for the drift.

Euler-Maruyama scheme of the limiting SDE. Denote $\Delta_L = \frac{1}{L}$, $\{t_k = k/L, \, k = 0, 1, \ldots, L\}$ and $\Delta V_k^A = V_{t_{k+1}}^A - V_{t_k}^A$ and $\Delta V_k^b = V_{t_{k+1}}^b - V_{t_k}^b$. The Euler-Maruyama discretization of the SDE (26) is defined as:

$$\widehat{h}_{k+1}^{(L)} - \widehat{h}_k^{(L)} = \mu\left(t_k, \widehat{h}_k^{(L)}\right) \Delta_L + \Delta V_k^A \, \widehat{h}_k^{(L)} + \Delta V_k^b = \widehat{h}_k^{(L)} + f^{(L)}\left(k, \widehat{h}_k^{(L)}\right) \tag{28}$$

where

$$f^{(L)}(k,h) = \mu\left(t_k, h\right) \Delta_L + \Delta V_k^A \, h + \Delta V_k^b. \tag{29}$$

Define the *continuous-time extension* of the hidden state dynamics

$$\overline{H}_t^{(L)} = h_k^{(L)} \mathbf{1}_{t_k \leq t < t_{k+1}}, \qquad k = 0, \ldots, L-1 \tag{30}$$

and denote

$$\begin{aligned} M_k^{(L)}(h) &= \left(\mu\left(t_k, h\right) - \frac{1}{2}\sigma''(0) Q(t_k, h) \right) \Delta_L + \Delta V_k^A \, h + \Delta V_k^b \\ &= \left(U_{t_k}^A \, h + \, U_{t_k}^b + \overline{A}_{t_k} h + \bar{b}_{t_k} \right) \Delta_L + \Delta V_k^A \, h + \Delta V_k^b \\ &=: \ \widetilde{\mu}\left(t_k, h\right) \Delta_L + \Delta V_k^A \, h + \Delta V_k^b, \end{aligned}$$

From (13) we thus have

$$\Delta h_k^{(L)} := h_{k+1}^{(L)} - h_k^{(L)} = \sigma\left(M_k^{(L)}(h_k^{(L)})\right).$$

For any vector $x \in \mathbb{R}^d$, denote $(x)_i$ as the i-th component of x ($i = 1, 2, \ldots, d$). Further denote $\Delta h_k^{(L),i}$ and $M_k^{(L),i}$ the i-th element of $\Delta h_k^{(L)}$ and $M_k^{(L)}$, respectively. Applying a third-order Taylor expansion of σ around 0 with the help of Assumption 4.2, for $i = 1, 2, \ldots, d$, we get

$$\begin{aligned}
\Delta h_k^{(L),i} &= \sigma\left(M_k^{(L),i}(h_k^{(L)})\right) = M_k^{(L),i}(h_k^{(L)}) + \frac{1}{2}\sigma''(0)\left(M_k^{(L),i}(h_k^{(L)})\right)^2 + \frac{1}{6}\sigma'''(\nu_i)\left(M_k^{(L),i}(h_k^{(L)})\right)^3 \\
&= \underbrace{\mu_i\left(t_k, h_k^{(L)}\right)\Delta_L + (\Delta V_k^A\, h_k^{(L)})_i + (\Delta V_k^b)_i}_{f_i^{(L)}(k, h_k^{(L)})} + \underbrace{\frac{1}{2}\sigma''(0)\left(\left(M_k^{(L),i}(h_k^{(L)})\right)^2 - Q_i(t_k, h_k^{(L)})\right)}_{N_k^{(L),i}(h_k^{(L)})} + \frac{1}{6}\sigma'''(\nu_i)\left(M_k^{(L),i}(h_k^{(L)})\right)^3 \\
&= f_i^{(L)}(k, h_k^{(L)}) + N_k^{(L),i}(h_k^{(L)}) + \frac{1}{6}\sigma'''(\nu_i)\left(M_k^{(L),i}(h_k^{(L)})\right)^3,
\end{aligned}$$

with $|\nu_i| < \left|M_k^{(L),i}(h_k^{(L)})\right|$. The increment of the hidden state $\Delta h_k^{(L),i}$ has two parts: the increment of the Euler-Maruyama scheme $f_i^{(L)}(k, h_k^{(L)})$ and the residual $D_k^{(L),i}(h_k^{(L)}) := \frac{1}{6}\sigma'''(\nu_i)\left(M_k^{(L),i}(h_k^{(L)})\right)^3 + N_k^{(L),i}(h_k^{(L)})$. It is clear from here that the Euler-Maruyama scheme of the limiting SDE is different from the ResNet dynamics. Hence classical results on the convergence of discrete SDE schemes cannot be applied directly.

In our analysis it will be more natural to work with the following *continuous-time approximation*:

$$\widetilde{H}_t^{(L)} := h_0^{(L)} + \int_0^t \mu\left(t_{k_s}, \overline{H}_s^{(L)}\right) \mathrm{d}s + \int_0^t \left(\mathrm{d}V_s^A\, \overline{H}_s^{(L)} + \mathrm{d}V_s^b\right) + \sum_{k \le Lt} D_k^{(L)}\left(h_k^{(L)}\right), \tag{31}$$

where $D_k^{(L)}(h) = \left(D_k^{(L),1}(h), \ldots, D_k^{(L),d}(h)\right)^\top$ and k_s is the integer for which $s \in [t_{k_s}, t_{k_s+1})$ for a given $s \in [0, 1)$. From the above definitions we have $\widetilde{H}_{t_k}^{(L)} = \overline{H}_{t_k}^L = h_k^{(L)}$.

Lemma E.4 (Local Lipschitz condition and uniform integrability)**.** *Under the assumptions of Theorem 4.4,*

1. *For each $R > 0$, there exists a constant C_R, depending only on R, such that almost surely we have*

$$\|\mu(t, x) - \mu(t, y)\|^2 \le C_R \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d \text{ with } \|x\| \vee \|y\| \le R, \text{ and } \forall t \in [0, 1], \tag{32}$$

where μ is defined in (27).

2. *There exist constants $p > 2$ and $C > 0$ such that*

$$\mathbb{E}\left[\sup_{0 \le t \le 1} \left\|\widetilde{H}_t^{(L)}\right\|^p\right] \vee \mathbb{E}\left[\sup_{0 \le t \le 1} \|H_t\|^p\right] \le C. \tag{33}$$

Remark E.5. *Note that (Higham et al., 2002) assumes the uniform integrability condition for $\widetilde{H}_t^{(L)}$, which is difficult to verify in practice. Here we relax this condition by only assuming the uniform integrability condition for the ResNet dynamics $\{h_k^{(L)} : k = 0, \ldots, L\}$, see Assumption E.3. We can then prove (33) under Assumption E.3 and some properties of the Itô processes.*

Lemma E.4 is proved by first showing $Q(t, x)$ is locally Lipschiz and then by applying Minkowski inequality to $\left\|\overline{H}_s^{(L)} - \widetilde{H}_s^{(L)}\right\|^p$ with $p = \frac{1}{2}p_1 > 2$, where p_1 is defined in (23).

We are now ready to prove Theorem 4.4.

Proof. Let us define two stopping times to utilize the local Lipschitz property of μ:

$$\tau_R := \inf\left\{t \ge 0 : \left\|\widetilde{H}_t^{(L)}\right\| \ge R\right\}, \quad \rho_R := \inf\left\{t \ge 0 : \|H_t\| \ge R\right\}, \quad \theta_R := \tau_R \wedge \rho_R, \tag{34}$$

and define the approximation errors

$$e_1(t) := \widetilde{H}_t^{(L)} - H_t, \text{ and } e_2(t) := \widetilde{H}_t^{(L)} - \overline{H}_t^{(L)}. \tag{35}$$

The proof contains two steps. The first step is to show $\lim_{L\to\infty} \mathbb{E}\left[\sup_{0\le t\le 1} \|e_1(t)\|^2\right] = 0$ and the second step is to show $\lim_{L\to\infty} \mathbb{E}\left[\sup_{0\le t\le 1} \|e_2(t)\|^2\right] = 0$.

Step 1: $\widetilde{H}$ and H are uniformly close to each other. Following the idea in (Higham et al., 2002), we first show that for any $\delta > 0$ (to be determined later), by Young's inequality,

$$\mathbb{E}\left[\sup_{0\le t\le 1} \|e_1(t)\|^2\right] \le \mathbb{E}\left[\sup_{0\le t\le 1} \left\|\widetilde{H}_{t\wedge\theta_R}^{(L)} - H_{t\wedge\theta_R}\right\|^2\right] + \frac{2^{p+1}\delta C}{p} + \frac{(p-2)2C}{p\delta^{2/(p-2)}R^p}, \tag{36}$$

where C and p are defined in (33). Now, we bound the first term on the right-hand side of (36). Using the definition of the targeted SDE limit in (26), the continuous-time approximation (31), and the Cauchy-Schwarz inequality, we get

$$\begin{aligned}\left\|\widetilde{H}_{t\wedge\theta_R}^{(L)} - H_{t\wedge\theta_R}\right\|^2 \le\ & 4\left[\int_0^{t\wedge\theta_R} \left\|\mu\left(s, \overline{H}_s^{(L)}\right)\mathrm{d}s - \mu\left(s, H_s\right)\right\|^2 \mathrm{d}s\right] + 4\left[\int_0^{t\wedge\theta_R} \left\|\mu\left(t_{k_s}, \overline{H}_s^{(L)}\right)\mathrm{d}s - \mu\left(s, \overline{H}_s^{(L)}\right)\right\|^2 \mathrm{d}s\right] \\ & +4\left\|\int_0^{t\wedge\theta_R} \mathrm{d}W_s^A\left(\overline{H}_s^{(L)} - H_s\right)\right\|^2 + 4\left\|\sum_{k\le L(t\wedge\theta_R)} D_k^{(L)}\left(h_k^{(L)}\right)\right\|^2.\end{aligned}$$

Therefore, from the local Lipschitz condition (32) and Doob's martingale inequality (Revuz & Yor, 2013), we have for any $\tau \le 1$,

$$\begin{aligned}&\mathbb{E}\left[\sup_{0\le t\le\tau}\left\|\widetilde{H}_{t\wedge\theta_R}^{(L)} - H_{t\wedge\theta_R}\right\|^2\right] \\ \le\ & 32\left(C_R + C_1^2\right)\int_0^\tau \mathbb{E}\left[\sup_{0\le r\le s}\left\|\widetilde{H}_{r\wedge\theta_R}^{(L)} - H_{r\wedge\theta_R}\right\|^2\right]\mathrm{d}s + 32\left(C_R + C_1^2\right)\underbrace{\mathbb{E}\int_0^{\tau\wedge\theta_R}\left\|\overline{H}_s^{(L)} - \widetilde{H}_s^{(L)}\right\|^2\mathrm{d}s}_{\textcircled{1}} \\ & +4\underbrace{\mathbb{E}\left[\int_0^{t\wedge\theta_R}\left\|\mu\left(t_{k_s}, \overline{H}_s^{(L)}\right) - \mu\left(s, \overline{H}_s^{(L)}\right)\right\|^2\mathrm{d}s\right]}_{\textcircled{2}} + 4\underbrace{\mathbb{E}\left[\sup_{0\le t\le\tau}\left\|\sum_{k\le L(t\wedge\theta_R)} D_k^{(L)}\left(h_k^{(L)}\right)\right\|^2\right]}_{\textcircled{3}}.\end{aligned} \tag{37}$$

Upper bound for ②. By the Cauchy–Schwarz inequality, the following holds for almost all $h \in \mathbb{R}^d$:

$$\|\mu(t,h) - \mu(s,h)\|^2 \le C_M |t-s|^\kappa \left(1 + \|h\|^2 + \|h\|^4\right). \tag{38}$$

Under Assumption E.3, there exists a constant $\widetilde{C}_0 > 0$ such that

$$\mathbb{E}\left[\sup_{0\le t\le 1}\left(\left\|\overline{H}_t^{(L)}\right\|^4 + \left\|\overline{H}_t^{(L)}\right\|^2\right)\right] \le \widetilde{C}_0. \tag{39}$$

Hence by Tonelli's theorem,

$$\begin{aligned}&\mathbb{E}\left[\int_0^{t\wedge\theta_R}\left\|\mu\left(t_{k_s}, \overline{H}_s^{(L)}\right) - \mu\left(s, \overline{H}_s^{(L)}\right)\right\|^2\mathrm{d}s\right] \le \int_0^1 \mathbb{E}\left[\left\|\mu\left(t_{k_s}, \overline{H}_s^{(L)}\right) - \mu\left(s, \overline{H}_s^{(L)}\right)\right\|^2\right]\mathrm{d}s \\ \le\ & (\widetilde{C}_0 + 1)C_M L\left(\int_0^{1/L} r^\kappa \mathrm{d}r\right) = \frac{(\widetilde{C}_0 + 1)C_M}{1+\kappa}L^{-\kappa}.\end{aligned} \tag{40}$$

Upper bound for ③. Define the following discrete filtration $\mathcal{G}_k := \sigma\Big(U^A_s, U^A_s, q^A_s, q^b_s, B^A_s, B^b_s \ : \ s \le t_{k+1}\Big)$. Note that $h^{(L)}_k$ is $\mathcal{G}_{k-1}$-measurable but not $\mathcal{G}_k$-measurable. Define for $k = 0, \dots, L-1$ and for $i = 1, \dots, d$:

$$\begin{aligned}
X^i_k &:= \left(\left(\Delta V^A_k\, h^{(L)}_k\right)_i + \left(\Delta V^b_k\right)_i\right)^2 - \mathbb{E}\left[\left(\left(\Delta V^A_k\, h^{(L)}_k\right)_i + \left(\Delta V^b_k\right)_i\right)^2 \middle| \mathcal{G}_{k-1}\right] \\
Y^i_k &:= \mathbb{E}\left[\left(\left(\Delta V^A_k\, h^{(L)}_k\right)_i + \left(\Delta V^b_k\right)_i\right)^2 \middle| \mathcal{G}_{k-1}\right] - Q_i\left(t_k, h^{(L)}_k\right)\Delta_L \\
J^i_k &:= \widetilde{\mu}_i(t,h)^2(\Delta_L)^2 + 2\widetilde{\mu}_i(t,h)\Delta_L\left(\left(\Delta V^A_k\, h\right)_i + \left(\Delta V^b_k\right)_i\right).
\end{aligned}$$

We can then deduce the following bound on ③ by Cauchy-Schwarz.

$$\begin{aligned}
&\mathbb{E}\left[\sup_{0\le t\le\tau}\left|\sum_{k\le L(t\wedge\theta_R)} D^{(L),i}_k\left(h^{(L)}_k\right)\right|^2\right] \\
\le\ & \sigma''(0)^2\,\mathbb{E}\left[\sup_{0\le t\le\tau}\left|\sum_{k\le Lt} X^i_k \mathbb{1}_{\left\|h^{(L)}_k\right\|\le R}\right|^2\right] + \sigma''(0)^2\,\mathbb{E}\left[\sup_{0\le t\le\tau}\left|\sum_{k\le Lt} Y^i_k \mathbb{1}_{\left\|h^{(L)}_k\right\|\le R}\right|^2\right] \\
&+ \sigma''(0)^2 L\sum_{k=0}^{L-1}\mathbb{E}\left[\left|J^i_k\right|^2\mathbb{1}_{\left\|h^{(L)}_k\right\|\le R}\right] + \frac{(\sigma'''(\nu_i))^2}{9}L\sum_{k=0}^{L-1}\mathbb{E}\left[\left|M^{(L),i}_k\left(h^{(L)}_k\right)\right|^6\mathbb{1}_{\left\|h^{(L)}_k\right\|\le R}\right] \qquad (41)
\end{aligned}$$

We provide an upper bound for each of the four terms in (41). For the first term, denote $\widetilde{X}^i_k = X^i_k\mathbb{1}\left(\left\|h^{(L)}_k\right\| \le R\right)$ and $S^i_k = \sum_{k'=0}^k \widetilde{X}^i_{k'}$ so that $\left(S^i_k\right)_{k=-1,0,\dots,L-1}$ is a $(\mathcal{G}_k)$-martingale. Hence, by Doob's martingale inequality, we have

$$\mathbb{E}\left[\sup_{0\le t\le\tau}\left|\sum_{k\le Lt} X^i_k\mathbb{1}_{\left\|h^{(L)}_k\right\|\le R}\right|^2\right] = \mathbb{E}\left[\sup_{0\le t\le\tau}\left|S^i_{\lfloor Lt\rfloor}\right|^2\right] \le 4\,\mathbb{E}\left[\left|S^i_{\lfloor L\tau\rfloor}\right|^2\right]. \qquad (42)$$

Fix $k = 0, \dots, L-1$. For every $i = 1, \dots, d$, we compute the following conditional expectation.

$$\mathbb{E}\left[\left(S^i_k\right)^2 \;\middle|\; \mathcal{G}_{k-1}\right] = \mathbb{E}\left[\left(S^i_{k-1}\right)^2 + 2\widetilde{X}^i_k\sum_{k'=0}^{k-1}\widetilde{X}^i_{k'} + \left(\widetilde{X}^i_k\right)^2 \;\middle|\; \mathcal{G}_{k-1}\right] = \left(S^i_{k-1}\right)^2 + \mathbb{E}\left[\left(\widetilde{X}^i_k\right)^2 \;\middle|\; \mathcal{G}_{k-1}\right] \qquad (43)$$

The cross-term disappear as $\mathbb{E}\left[\widetilde{X}^i_k \;\middle|\; \mathcal{G}_{k-1}\right] = \mathbb{E}\left[X^i_k \mid \mathcal{G}_{k-1}\right]\mathbb{1}\left(\left\|h^{(L)}_k\right\| \le R\right) = 0$ by definition of X^i_k. Furthermore, conditionally on $\mathcal{G}_{k-1}$ and on $\left\{\left\|h^{(L)}_k\right\| \le R\right\}$, observe that X^i_k is the centered square of a normal random variable whose variance is $\mathcal{O}(L^{-1})$ uniformly in k by (19), so there exist $C_{R,1} > 0$ depending only on R such that

$$\sup_{k=0,\dots,K-1}\mathbb{E}\left[\left(\widetilde{X}^i_k\right)^2 \;\middle|\; \mathcal{G}_{k-1}\right] \le C_{R,1}L^{-2}.$$

Hence, plugging back into (42), we obtain

$$\mathbb{E}\left[\sup_{0\le t\le\tau}\left|\sum_{k\le Lt} X^i_k\mathbb{1}_{\left\|h^{(L)}_k\right\|\le R}\right|^2\right] \le 4C_{R,1}L^{-1}. \qquad (44)$$

For the second term involving Y^i_k, we explicitly compute the conditional expectation using the definition of V in (27) and the definition of Q in (12).

$$Y^i_k = \int_{t_k}^{t_{k+1}}\left(\mathbb{E}\left[\Sigma^b_{s,ii} - \Sigma^b_{t_k,ii} \mid \mathcal{G}_{k-1}\right] + \sum_{j,l=1}^d h^{(L)}_{k,j}h^{(L)}_{k,l}\,\mathbb{E}\left[\Sigma^A_{s,ijil} - \Sigma^A_{t_k,ijil} \mid \mathcal{G}_{k-1}\right]\right)\mathrm{d}s.$$

Now, we compute directly the following bound by Cauchy-Schwarz, Tonelli and (20) in Assumption E.1-(ii):

$$\mathbb{E}\left[\sup_{0\le t\le\tau}\left|\sum_{k\le L(t\wedge\theta_R)} Y_k^i \mathbb{1}_{\left\|h_k^{(L)}\right\|\le R}\right|^2\right] \le M(1+R^2)^2\left(L\int_0^{1/L} r^{\kappa/2}\mathrm{d}r\right)^2 =: C_{R,2}L^{-\kappa}, \tag{45}$$

where $C_{R,2} > 0$ depends only on R. Moving to the third term of (41) involving J_k^i, we get directly from the definition of V^A and V^b that there exists $C_{R,3} > 0$ depending only on R such that

$$\sup_{\|h\|\le R}\mathbb{E}\left[\left|J_k^i\right|^2\mathbb{1}_{\left\|h_k^{(L)}\right\|\le R}\right] \le C_{R,3}L^{-3}. \tag{46}$$

Finally, we bound the fourth term of (41) using Cauchy-Schwarz, Assumption 4.2 and property (19) of the Itô processes:

$$\sigma'''(\nu_i)^2\sup_{\|h\|\le R}\mathbb{E}\left[\left(M_k^{(L),i}(h)\right)^6\right] \le m^2\, C_{R,4}L^{-3}, \tag{47}$$

for some constant $C_{R,4} > 0$ depending only on R. Combining the results in (44), (45), (46) and (47), we have

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le t\le\tau}\left|\sum_{k\le L(t\wedge\theta_R)} D_k^{(L),i}\left(h_k^{(L)}\right)\right|^2\right] &\le 4\sigma''(0)^2C_{R,1}L^{-1} + \sigma''(0)^2C_{R,2}L^{-\kappa} + \sigma''(0)^2C_{R,3}L^{-1} + \frac{m^2}{9}C_{R,4}L^{-1}\\ &=: \frac{C_{R,5}}{4d}L^{-\kappa} + \frac{C_{R,6}}{4d}L^{-1}.\end{aligned} \tag{48}$$

Upper bound for ①. Given $s \in [0, T\wedge\theta_R)$, denote k_s as the integer for which $s \in [t_k, t_{k_s+1})$. Then

$$\begin{aligned}\overline{H}_s^{(L)} - \widetilde{H}_s^{(L)} &= h_{k_s}^{(L)} - \left(h_{k_s}^{(L)} + \int_{t_{k_s}}^s \mu(s,\overline{H}_s^{(L)})\mathrm{d}s + \int_{t_{k_s}}^s\left(\mathrm{d}V_s^A\overline{H}_s^{(L)} + \mathrm{d}V_s^b\right)\right)\\ &= -\mu\left(t_{k_s}, h_{k_s}^{(L)}\right)(s-t_{k_s}) - \left(V_s^A - V_{t_{k_s}}^A\right)h_{k_s}^{(L)} - \left(V_s^b - V_{t_{k_s}}^b\right),\end{aligned} \tag{49}$$

and by the Mean-value Theorem and the continuity of μ. Hence

$$\left\|\overline{H}_s^{(L)} - \widetilde{H}_s^{(L)}\right\|^2 \le 3\left\|\mu\left(t_{k_s}, h_{k_s}^{(L)}\right)\right\|^2(\Delta_L)^2 + 3\left\|h_{k_s}^{(L)}\right\|^2\left\|V_s^A - V_{t_{k_s}}^A\right\|^2 + 3\left\|V_s^b - V_{t_{k_s}}^b\right\|^2. \tag{50}$$

Now, from the local Lipschitz condition (32), for $\|h\| \le R$ we have almost surely

$$\|\mu(s,h)\|^2 \le 2\left(\|\mu(s,h)-\mu(s,0)\|^2 + \|\mu(s,0)\|^2\right) \le 2\left(C_R\|h\|^2 + \|\mu(s,0)\|^2\right).$$

Hence,

$$(50) \le 4\left(C_R\left\|h_{k_s}^{(L)}\right\|^2 + \|\mu(s,0)\|^2 + 1\right)\left(\Delta_L^2 + \left\|V_s^A - V_{t_{k_s}}^A\right\|^2 + \left\|V_s^b - V_{t_{k_s}}^b\right\|^2\right).$$

Hence, using (33) and the Lyapunov inequality (Platen & Bruti-Liberati, 2010), we get

$$\mathbb{E}\int_0^{\tau\wedge\theta_R}\left\|\overline{H}_s^{(L)} - \widetilde{H}_s^{(L)}\right\|^2\mathrm{d}s \le 4\left(C_R\,C_0^{2/p} + 1 + \int_0^1\|\mu(s,0)\|^2\mathrm{d}s\right)\left(\Delta_L^2 + 4C_1\Delta_L\right). \tag{51}$$

Combining everything: From (40), (48) and (51), we have in (37) that

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le t\le\tau}\left\|\widetilde{H}_{\tau\wedge\theta_R}^{(L)} - H_{t\wedge\theta_R}\right\|^2\right] &\le 128(C_R + C_1^2)\left(C_R\,C_0^{2/p} + 1 + \int_0^1\|\mu(s,0)\|^2\mathrm{d}s\right)\left(L^{-2} + 4C_1L^{-1}\right)\\ &+ \frac{(\widetilde{C}_0+1)C_M}{1+\kappa}L^{-\kappa} + \left(C_{R,5}L^{-\kappa} + C_{R,6}L^{-1}\right) + 32(C_R + C_1^2)\int_0^\tau\mathbb{E}\left[\sup_{0\le r\le s}\left\|\widetilde{H}_{r\wedge\theta_R}^{(L)} - H_{r\wedge\theta_R}\right\|^2\right]\mathrm{d}s.\end{aligned}$$

Applying the Grönwall inequality,

$$\mathbb{E}\left[\sup_{0\le t\le \tau}\left\|\widetilde{H}^{(L)}_{\tau\wedge\theta_R}-H_{t\wedge\theta_R}\right\|^2\right]\le C_9L^{-\min\{1,\kappa\}}\Big(C_R^2+C_{R,5}+C_{R,6}+1\Big)e^{32(C_R+C_1^2)}, \tag{52}$$

where C_9 is a universal constant independent of L, R and δ. Combining (52) with (36), we have

$$\mathbb{E}\left[\sup_{0\le t\le 1}\|e_1(t)\|^2\right]\le C_9L^{-\min\{1,\kappa\}}\Big(C_R^2+C_{R,5}+C_{R,6}+1\Big)e^{32(C_R+C_1^2)}+\frac{2^{p+1}\delta C}{p}+\frac{(p-2)2C}{p\delta^{2/(p-2)}R^p}. \tag{53}$$

Given any $\epsilon>0$, we can choose $\delta>0$ so that $\frac{2^{p+1}\delta C}{p}<\frac{\epsilon}{3}$, then choose R so that $\frac{(p-2)2C}{p\delta^{2/(p-2)}R^p}<\frac{\epsilon}{3}$, and finally choose L sufficiently large so that

$$C_9L^{-\min\{1,\kappa\}}\Big(C_R^2+C_{R,5}+C_{R,6}+1\Big)e^{32(C_R+C_1^2)}\le\frac{\epsilon}{3}.$$

Therefore in (53), we have,

$$\mathbb{E}\left[\sup_{0\le t\le 1}\|e_1(t)\|^2\right]\le\epsilon. \tag{54}$$

Step 2: $\overline{H}$ and $\widetilde{H}$ are uniformly close to each other. Recall the relationship between $\widetilde{H}$ and $\overline{H}$ defined in (49): by (19) we have almost surely that

$$\left\|\overline{H}^{(L)}_s-\widetilde{H}^{(L)}_s\right\|^2\le C_{10}\left(\left\|h^{(L)}_{k_s}\right\|^4+\left\|h^{(L)}_{k_s}\right\|^2+1\right)(\Delta_L)^2+3\left(\left\|h^{(L)}_{k_s}\right\|^2\left\|V^A_s-V^A_{t_{k_s}}\right\|^2+\left\|V^b_s-V^b_{t_{k_s}}\right\|^2\right).$$

Therefore

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le s\le 1}\left\|\overline{H}^{(L)}_s-\widetilde{H}^{(L)}_s\right\|^2\right]&\le C_{10}\left(\mathbb{E}\left[\sup_{0\le s\le 1}\left\|h^{(L)}_{k_s}\right\|^4\right]+\mathbb{E}\left[\sup_{0\le s\le 1}\left\|h^{(L)}_{k_s}\right\|^2\right]+1\right)(\Delta_L)^2\\&+3\left(\left(\mathbb{E}\left[\sup_{0\le s\le 1}\left\|h^{(L)}_{k_s}\right\|^4\right]\mathbb{E}\left[\sup_{0\le s\le 1}\left\|V^A_s-V^A_{t_{k_s}}\right\|^4\right]\right)^{1/2}+\mathbb{E}\left[\sup_{0\le s\le 1}\left\|V^b_s-V^b_{t_{k_s}}\right\|^2\right]\right).\end{aligned} \tag{55}$$

By the Power Mean inequality and Doob's martingale inequality,

$$\mathbb{E}\left[\sup_{0\le s\le 1}\left\|V^A_s-V^A_{t_{k_s}}\right\|^4\right]\le\mathbb{E}\left[\sum_{k=0}^{L-1}\left(\sup_{t_k\le s<t_{k+1}}\left\|V^A_s-V^A_{t_{k_s}}\right\|^4\right)\right]\le C_{11}\Delta_L. \tag{56}$$

Using Hölder's inequality yields

$$\mathbb{E}\left[\sup_{0\le s\le 1}\left\|V^A_s-V^A_{t_{k_s}}\right\|^2\right]\le\left(\mathbb{E}\left[\sup_{0\le s\le 1}\left\|V^A_s-V^A_{t_{k_s}}\right\|^4\right]\right)^{1/2}\le\sqrt{C_{11}}\Delta_L^{1/2}. \tag{57}$$

Combining (39), (56), and (57) in (55), we obtain

$$\mathbb{E}\left[\sup_{0\le t\le 1}\|e_2(t)\|^2\right]=\mathbb{E}\left[\sup_{0\le t\le 1}\left\|\overline{H}^{(L)}_t-\widetilde{H}^{(L)}_t\right\|^2\right]\le C_{12}\Delta_L^{1/2},$$

for some constant $C_{12}>0$. By choosing $L>(C_{12}/\epsilon)^2$, we have

$$\mathbb{E}\left[\sup_{0\le t\le 1}\|e_2(t)\|^2\right]\le\epsilon. \tag{58}$$

Finally, combining (54) and (58) leads to the desired result.

□