Appendix

A0 An overview of signal propagation in wide neural networks

In section, we review some results and tools from the theory of signal of propagation in wide neural networks. This will prove valuable in the rest of the appendix.

A0.1 Neural Network Gaussian Process (NNGP)

Standard ResNet without $\mathcal{SD}$. Consider a standard ResNet architecture with L layers. The forward propagation of some input $x \in \mathbb{R}^d$ is given by

$$\begin{aligned} y_0(x) &= \Psi_0(x, W_0) \\ y_l(x) &= y_{l-1}(x) + \Psi_l(y_{l-1}(x), W_l), \quad 1 \leq l \leq L, \\ y_{out}(x) &= \Psi_{out}(y_L(x), W_{out}), \end{aligned} \tag{A1}$$

where W_l are the weights in the l^{th} layer, Ψ is a mapping that defines the nature of the layer, and y_l are the pre-activations. we consider constant width ResNet and we further denote by N the width, i.e. for all $l \in [L-1]$, $y_l \in \mathbb{R}^N$. The output function of the network is given by $s(y_{out})$ where s is some convenient mapping for the learning task, e.g. the Softmax mapping for classification tasks. We denote by o the dimension of the network output, i.e. $s(y_{out}) \in \mathbb{R}^o$ which is also the dimension of y_{out}. For our theoretical analysis, we consider residual blocks composed of a Fully Connected linear layer

$$\Psi_l(x, W) = W\phi(x).$$

where $\phi(x)$ is the activation function. The weights are initialized with He init He et al. [2015], e.g. for ReLU, $W_{ij}^l \sim \mathcal{N}(0, 2/N)$.

The neurons $\{y_0^i(x)\}_{i\in[1:N]}$ are iid normally distributed random variables since the weights connecting them to the inputs are iid normally distributed. Using the Central Limit Theorem, as $N_0 \to \infty$, $y_1^i(x)$ becomes a Gaussian variable for any input x and index $i \in [1:N]$. Additionally, the variables $\{y_1^i(x)\}_{i\in[1:N]}$ are iid. Thus, the processes $y_1^i(.)$ can be seen as independent (across i) centred Gaussian processes with covariance kernel Q_1. This is an idealized version of the true process corresponding to letting width $N_0 \to \infty$. Doing this recursively over l leads to similar results for $y_l^i(.)$ where $l \in [1:L]$, and we write accordingly $y_l^i \overset{ind}{\sim} \mathcal{GP}(0, Q_l)$. The approximation of $y_l^i(.)$ with a Gaussian process was first proposed by [Neal, 1995] for single layer FeedForward neural networks and was extended recently to multiple feedforward layers by [Lee et al., 2019] and [Matthews et al., 2018]. More recently, excellent work by [Yang, 2020] introduced a unifying framework named Tensor Programs, confirming the large-width Gaussian Process behaviour for nearly all neural network architectures.

For any input $x \in \mathbb{R}^d$, we have $\mathbb{E}[y_l^i(x)] = 0$, so that the covariance kernel is given by $Q_l(x, x') = \mathbb{E}[y_l^1(x)y_l^1(x')]$. It is possible to evaluate the covariance kernels layer by layer, recursively. More precisely, assume that y_{l-1}^i is a Gaussian process for all i. Let $x, x' \in \mathbb{R}^d$. We have that

$$\begin{aligned} Q_l(x, x') &= \mathbb{E}[y_l^1(x)y_l^1(x')] \\ &= \mathbb{E}[y_{l-1}^1(x)y_{l-1}^1(x')] + \sum_{j=1}^{N_{l-1}} \mathbb{E}[(W_l^{1j})^2\phi(y_{l-1}^j(x))\phi(y_{l-1}^j(x'))] \\ &\quad + \mathbb{E}\left[\sum_{j=1}^{N_{l-1}} W_l^{1j}(y_{l-1}^1(x)\phi(y_{l-1}^1(x')) + y_{l-1}^1(x')\phi(y_{l-1}^1(x)))\right]. \end{aligned}$$

Some terms vanish because $\mathbb{E}[W_l^{1j}] = 0$. Let $Z_j = \sqrt{\frac{N}{2}}W_l^{1j}$. The second term can be written as

$$\mathbb{E}\left[\frac{2}{N}\sum_j (Z_j)^2\phi(y_{l-1}^j(x))\phi(y_{l-1}^j(x'))\right] \to 2\,\mathbb{E}[\phi(y_{l-1}^1(x))\phi(y_{l-1}^1(x'))]\,,$$

where we have used the Central Limit Theorem. Therefore, the kernel Q_l satisfies for all $x, x' \in \mathbb{R}^d$

$$Q_l(x, x') = Q_{l-1}(x, x') + \mathcal{F}_{l-1}(x, x')\,,$$

where $\mathcal{F}_{l-1}(x, x') = 2\,\mathbb{E}[\phi(y^1_{l-1}(x))\phi(y^1_{l-1}(x'))]$.

For the ReLU activation function $\phi : x \mapsto \max(0, x)$, the recurrence can be written more explicitly as in [Hayou et al., 2019]. Let C_l be the correlation kernel, defined as

$$C_l(x, x') = \frac{Q_l(x,x')}{\sqrt{Q_l(x,x)Q_l(x',x')}} \tag{A2}$$

and let $f : [-1, 1] \to \mathbb{R}$ be given by

$$f : \gamma \mapsto \tfrac{1}{\pi}(\sqrt{1-\gamma^2} + \gamma \arcsin \gamma) + \frac{1}{2}\gamma\,. \tag{A3}$$

The recurrence relation reads

$$\begin{aligned} Q_l &= Q_{l-1} + \tfrac{f(C_{l-1})}{C_{l-1}} Q_{l-1}\,, \\ Q_0(x, x') &= 2\,\tfrac{x\cdot x'}{d}\,. \end{aligned} \tag{A4}$$

Standard ResNet with $\mathcal{SD}$. The introduction of the binary mask $\boldsymbol{\delta}$ in front of the residual blocks slightly changes the recursive expression of the kernel Q_l. It ii easy to see that with $\mathcal{SD}$, the Q_l follows

$$\begin{aligned} Q_l &= Q_{l-1} + p_l \tfrac{f(C_{l-1})}{C_{l-1}} Q_{l-1}\,, \\ Q_0(x, x') &= 2\,\tfrac{x\cdot x'}{d}\,. \end{aligned} \tag{A5}$$

where f is given by Eq. (A3).

We obtain similar formulas for Stable ResNet with and without $\mathcal{SD}$.

Stable ResNet without $\mathcal{SD}$.

$$\begin{aligned} Q_l &= Q_{l-1} + \frac{1}{L} \tfrac{f(C_{l-1})}{C_{l-1}} Q_{l-1}\,, \\ Q_0(x, x') &= 2\,\tfrac{x\cdot x'}{d}\,. \end{aligned} \tag{A6}$$

Stable ResNet with $\mathcal{SD}$.

$$\begin{aligned} Q_l &= Q_{l-1} + \frac{p_l}{L} \tfrac{f(C_{l-1})}{C_{l-1}} Q_{l-1}\,, \\ Q_0(x, x') &= 2\,\tfrac{x\cdot x'}{d}\,. \end{aligned} \tag{A7}$$

A0.2 Diagonal elements of the kernel Q_l

Lemma A1 (Diagonal elements of the covariance)**.** *Consider a ResNet of the form Eq. (1) (standard ResNet) or Eq. (3)(Stable ResNet), and let $x \in \mathbb{R}^d$. We have that for all $l \in [1 : L]$,*

- *Standard ResNet without $\mathcal{SD}$: $Q_l(x, x) = 2^l Q_0(x, x)$.*
- *Standard ResNet with $\mathcal{SD}$: $Q_l(x, x) = \prod_{k=1}^{l}(1 + p_k) Q_0(x, x)$.*
- *Stable ResNet without $\mathcal{SD}$: $Q_l(x, x) = (1 + \frac{1}{L})^l Q_0(x, x)$.*
- *Stable ResNet with $\mathcal{SD}$: $Q_l(x, x) = \prod_{k=1}^{l}(1 + \frac{p_k}{L}) Q_0(x, x)$.*

Proof. Let us prove the result for Standard ResNet with $\mathcal{SD}$. The proof is similar for the other cases. Let $x \in \mathbb{R}^d$. We know that

$$Q_l(x, x) = Q_{l-1}(x, x) + p_l f(1) Q_{l-1}(x, x)\,,$$

where f is given by Eq. (A3). It is straightforward that $f(1) = 1$. This yields

$$Q_l(x, x) = (1 + p_l) Q_{l-1}(x, x)\,.$$

we conclude by telescopic product. □

A0.3 Assumption 1 and gradient backpropagation

For gradient back-propagation, an essential assumption in the literature on signal propagation analysis in deep neural networks is that of the gradient independence which is similar in nature to the practice of feedback alignment [Lillicrap et al., 2016]. This assumption (Assumption 1) allows for derivation of recursive formulas for gradient back-propagation, and it has been extensively used in literature and empirically verified; see references below.

Gradient Covariance back-propagation. Assumption 1 was used to derive analytical formulas for gradient covariance back-propagation in a stream of papers; to cite a few, [Hayou et al., 2019, **?**, Poole et al., 2016, Xiao et al., 2018, Yang and Schoenholz, 2017]. It was validated empirically through extensive simulations that it is an excellent tool for FeedForward neural networks in Schoenholz et al. [2017], for ResNets in Yang and Schoenholz [2017] and for CNN in Xiao et al. [2018].

Neural Tangent Kernel (NTK). Assumption 1 was implicitly used by Jacot et al. [2018] to derive the recursive formula of the infinite width Neural Tangent Kernel (See Jacot et al. [2018], Appendix A.1). Authors have found that this assumption yields excellent empirical match with the exact NTK. It was also used later in [Arora et al., 2019, Hayou et al., 2020] to derive the infinite depth NTK for different architectures.

When used for the computation of gradient covariance and Neural Tangent Kernel, [Yang, 2020] proved that Assumption 1 yields the exact computation of the gradient covariance and the NTK in the limit of infinite width. We state the result for the gradient covariance formally.

Lemma A2 (Corollary of Theorem D.1. in [Yang, 2020])**.** *Consider a ResNet of the form (1) or (3) with weights W. In the limit of infinite width, we can assume that W^T used in back-propagation is independent from W used for forward propagation, for the calculation of Gradient Covariance.*

Lemma A3 (Gradient Second moment)**.** *Consider a ResNet of type Eq. (3) without $\mathcal{SD}$. Let $(x,t) \in \mathcal{D}$ be a sample from the dataset, and define the second of the gradient with $\tilde{q}_l(x,t) = \mathbb{E}_{\boldsymbol{W}}\left[\frac{\partial \mathcal{L}(x,\boldsymbol{W})}{\partial y_l}\right]$. Then, in the limit of infinite width, we have that*

$$\tilde{q}^l(x,t) = \left(1 + \frac{1}{L}\right) \tilde{q}^{l+1}(x,t)\,.$$

As a result, for all $l \in [1:L]$, we have that

$$\tilde{q}^l(x,t) = \left(1 + \frac{1}{L}\right)^{L-l} \tilde{q}^{L}(x,t)\,.$$

Proof. It is straighforward that

$$\frac{\partial \mathcal{L}(x,\boldsymbol{W})}{\partial y_l^i} = \frac{\partial \mathcal{L}(x,\boldsymbol{W})}{\partial y_{l+1}^i} + \frac{1}{\sqrt{L}} \sum_j \frac{\partial \mathcal{L}(x,\boldsymbol{W})}{\partial y_{l+1}^j} W_{l+1}^{ji} \phi'(y_l^i(x))\,.$$

Using lemma A2 and the Central Limit Theorem, we obtain

$$\tilde{q}^l(x,t) = \tilde{q}^{l+1}(x,t) + \frac{2}{L} \tilde{q}^{l+1}(x,t) \mathbb{E}[\phi'(y_l^1(x))^2]\,.$$

We conclude by observing that $\mathbb{E}[\phi'(y_i^l(x))^2] = \mathbb{P}(\mathcal{N}(0,1) > 0) = \frac{1}{2}$. □

A1 Proofs

A1.1 Proof of Lemma 1

Lemma 1 (Concentration of $L_{\boldsymbol{\delta}}$)**.** *For any $\beta \in (0,1)$, we have that with probability at least $1-\beta$,*

$$|L_{\boldsymbol{\delta}} - L_{\boldsymbol{p}}| \leq v_{\boldsymbol{p}}\, u^{-1}\left(\frac{\log(2/\beta)}{v_{\boldsymbol{p}}}\right) \tag{A8}$$

where $L_{\boldsymbol{p}} = \mathbb{E}[L_{\boldsymbol{\delta}}] = \sum_{l=1}^L p_l$, $v_{\boldsymbol{p}} = \text{Var}[L_{\boldsymbol{\delta}}] = \sum_{l=1}^L p_l(1-p_l)$, and $u(t) = (1+t)\log(1+t) - t$. Moreover, for a given average depth $L_{\boldsymbol{p}} = \bar{L}$, the upperbound in Eq. (2) is maximal for the uniform choice of survival probabilities $\boldsymbol{p} = \left(\frac{\bar{L}}{L}, ..., \frac{\bar{L}}{L}\right)$.

Proof. The concentration inequality is a simple application of Bennett's inequality: Let $X_1, ..., X_n$ be a sequence of independent random variables with finite variance and zero mean. Assume that there exists $a \in \mathbb{R}^+$ such that $X_i \leq a$ almost surely for all i. Define $S_n = \sum_{i=1}^n X_i$ and $\sigma_n^2 = \sum_{i=1}^n \mathbb{E}[X_i^2]$. Then, for any $t > 0$, we have that

$$\mathbb{P}(|S_n| > t) \leq 2 \exp\left(-\frac{\sigma_n^2}{a^2} u\left(\frac{at}{\sigma_n^2}\right)\right).$$

Now let us prove the second result. Fix some $\bar{L} \in (0, L)$. We start by proving that the function $\zeta(z) = z\, u^{-1}\left(\frac{\alpha}{z}\right)$ is increasing for any fixed $\alpha > 0$. Observe that $u'(t) = \log(1+t)$, so that

$$(u^{-1})'(z) = \frac{1}{u'(u^{-1}(z))} = \frac{1 + u^{-1}(z)}{z + u^{-1}(z)}.$$

This yields

$$\zeta'(z) = \frac{z u^{-1}\left(\frac{\alpha}{z}\right)^2 - \alpha}{\alpha + x u^{-1}\left(\frac{\alpha}{z}\right)}$$

For $z > 0$, the numerator is positive if and only if $\frac{\alpha}{z} > u\left(\sqrt{\frac{\alpha}{z}}\right)$, which is always true using the inequality $\log(1+t) < t$ for all $t > 0$.

Now let $\alpha = \log(2/\beta)$. Without restrictions on $L_{\boldsymbol{p}}$, it is straightforward that $v_{\boldsymbol{p}}$ is maximized by $\boldsymbol{p}' = (1/2, ..., 1/2)$. With the restriction $L_{\boldsymbol{p}} = \bar{L}$, the minimizer is the orthogonal projection of $\boldsymbol{p}'$ onto the convex set $\{\boldsymbol{p} : L_{\boldsymbol{p}} = \bar{L}\}$. This projection inherits the symmetry of $\boldsymbol{p}'$, which concludes the proof. □

A1.2 Proof of Proposition 1

Theorem 1 (Exploding gradient rate)**.** *Let $\mathcal{L}(x, z) = \ell(y_{out}(x; \boldsymbol{\delta}), z)$ for $(x, z) \in \mathbb{R}^d \times \mathbb{R}^o$, where $\ell(z, z')$ is some differentiable loss function. Let $\tilde{q}_l(x, z) = \mathbb{E}_{W,\boldsymbol{\delta}} \frac{\|\nabla_{y_l}\mathcal{L}\|^2}{\|\nabla_{y_L}\mathcal{L}\|^2}$, where the numerator and denominator are respectively the norms of the gradients with respect to the inputs of the l^{th} and L^{th} layers . Then, in the infinite width limit, under Assumption 1, for all $l \in [L]$ and $(x, z) \in \mathbb{R}^d \times \mathbb{R}^o$, we have*

- *With Stochastic Depth, $\tilde{q}_l(x, z) = \prod_{k=l+1}^{L}(1 + p_k)$*
- *Without Stochastic Depth (i.e. $\boldsymbol{\delta} = \mathbf{1}$), $\tilde{q}_l(x, z) = 2^{L-l}$*

Proof. It is straightforward that with Stochastic Depth

$$\frac{\partial \mathcal{L}(x, \boldsymbol{W})}{\partial y_l^i} = \frac{\partial \mathcal{L}(x, \boldsymbol{W})}{\partial y_{l+1}^i} + \delta_{l+1} \sum_j \frac{\partial \mathcal{L}(x, \boldsymbol{W})}{\partial y_{l+1}^j} W_{l+1}^{ji} \phi'(y_l^i(x)).$$

Denote for clarity for any neuron i and layer l, $d_{i,l} = \frac{\partial \mathcal{L}(x, \boldsymbol{W})}{\partial y_l^i}$. The equation can be written

$$d_{i,l} = d_{i,l+1} + \delta_{l+1} \sum_j d_{j,l+1} W_{l+1}^{ji} \phi'(y_l^i(x)) \tag{A9}$$

We notice that for any $k \geq l+1$ and any i, $d_{i,k}$ on depends on W_{l+1} through the forward pass, and that the terms W_{l+1}^{ji} in equation (A9) come from the backward pass. Therefore, Lemma A2 entails

$$\begin{aligned}
\mathbb{E}_{W_{l+1}^{\text{backward}}} \frac{\|\nabla_{y_l}\mathcal{L}\|^2}{\|\nabla_{y_L}\mathcal{L}\|^2} &= \frac{\sum_i d_{i,l+1}^2 + \delta_{l+1}^2 \frac{2}{N} \sum_i \phi'(y_l^i(x))^2 \sum_j d_{j,l+1}^2}{\sum_j d_{j,L}^2} \\
&= \frac{\|\nabla_{y_{l+1}}\mathcal{L}\|^2}{\|\nabla_{y_L}\mathcal{L}\|^2}(1 + \delta_{l+1} \frac{2}{N} \sum_i \phi'(y_l^i(x))^2),
\end{aligned}$$

where $\frac{2}{N}$ is the variance of W_{l+1}^{ji} (N is the width of the network). Again using Lemma A2 and taking the expectation with respect to the remaining weights and mask concludes the proof.

□

A1.3 Proof of Lemma 2

Lemma 2 (Maximal regularization)**.** *Consider the empirical loss $\mathcal{L}$ given by Eq. (6) for some fixed weights $\boldsymbol{W}$ (e.g. $\boldsymbol{W}$ could be the weights at any training step of SGD).Then, for a fixed training budget $\bar{L}$, the regularization is maximal for*

$$p_l^* = \min\left(1, \max(0, \frac{1}{2} - C g_l(\boldsymbol{W})^{-1})\right),$$

where C is a normalizing constant, that has the same sign as $L - 2\bar{L}$. The global maximum is obtained for $p_l = 1/2$.

Proof. Let $a_l = g_l(\boldsymbol{W})$. Noticing that $p_l(1-p_l) = 1/4 - (p_l - 1/2)^2$, it comes that

$$\begin{aligned}\sum_l p_l(1-p_l)a_l &= \frac{\sum a_l}{4} - \sum_l (p_l - 1/2)^2 a_l \\ &= \frac{\sum a_l}{4} - \left\| p \odot \sqrt{a} - \frac{\sqrt{a}}{2} \right\|_2^2,\end{aligned}$$

with the abuse of notations $\sqrt{a} = (\sqrt{a_1}, ..., \sqrt{a_L})$ and where $\odot$ stands for the element-wise product. We deduce from this expression that $p = \frac{1}{2}$ is the global maximizer of the regularization term. With a fixed training budget, notice that the expression is maximal for $p^* \odot \sqrt{a}$ the orthogonal projection of $\frac{\sqrt{a}}{2}$ on the intersection of the affine hyper-plane $\mathcal{H}$ containing the L points of the form $(0, ..., L_m\sqrt{a_l}, ..., 0)$ and the hyper-cube $\mathcal{C}$ of the points with coordinates in $[0, 1]$. Writing the KKT conditions yields for every l:

$$p_l = 1/2 - \beta a_l^{-1} - \lambda_{0,l}\mathbb{1}_{p_l=0} + \lambda_{1,l}\mathbb{1}_{p_l=1},$$

where $\lambda_{0,l}, \lambda_{1,l} \geq 0$. Taking $p = p^*$, $\beta = C$, $\lambda_{0,l} = 1/2 + \beta a_l^{-1}$ and $\lambda_{1,l} = 1/2 - \beta a_l^{-1}$. Since the program is quadratic, the KKT conditions are necessary and sufficient conditions, which concludes the proof.

□

A1.4 Proof of Theorem 1

Theorem 1 (p^* is uniform at initialization)**.** *Assume $\phi = ReLU$ and $\boldsymbol{W}$ are initialized with $\mathcal{N}(0, \frac{2}{N})$. Then, in the infinite width limit, under Assumption 1, for all $l \in [1 : L]$, we have*

$$\mathbb{E}_{\boldsymbol{W}}[g_l(\boldsymbol{W})] = \mathbb{E}_{\boldsymbol{W}}[g_1(\boldsymbol{W})].$$

As a result, given a budget $\bar{L}$, the average regularization term $\frac{1}{2L}\sum_{l=1}^{L} p_l(1-p_l)\mathbb{E}_{\boldsymbol{W}}[g_l(\boldsymbol{W})]$ is maximal for the uniform mode $\boldsymbol{p}^* = (\bar{L}/L, \ldots, \bar{L}/L)$.

Proof. Using the expression of g_l, we have that

$$\mathbb{E}_{\boldsymbol{W}}[g_l(\boldsymbol{W})] = \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{W}}[\|\zeta_l(x_i, \boldsymbol{W})\|_2^2].$$

Thus, to conclude, it is sufficient to show that for some arbitrary $i \in [1 : n]$, we have $\mathbb{E}_{\boldsymbol{W}}[\|\zeta_l(x_i, \boldsymbol{W})\|_2^2] = \mathbb{E}_{\boldsymbol{W}}[\|\zeta_1(x_i, \boldsymbol{W})\|_2^2]$ for all $l \in [1, L]$.
Fix $i \in [1 : n]$ and $l \in [1 : L]$. For the sake of simplicity, let $z_l = z_l(x; \mathbf{1})$ and $y_l = y_l(x; \mathbf{1})$. We have that

$$\|\zeta_l(x, \boldsymbol{W})\|_2^2 = \sum_{j=1}^{o} \langle \nabla_{y_l} G_l^j(y_l), z_l \rangle^2$$

Now let $j \in [1 : o]$. We have that

$$\langle \nabla_{y_l} G_l^j(y_l), z_l \rangle^2 = \sum_{k,k'} \frac{\partial G_l^j}{\partial y_l^k} z_l^k \frac{\partial G_l^j}{\partial y_l^{k'}} z_l^{k'}$$

Using Assumption 1, in the limit $N \to \infty$, we obtain

$$\mathbb{E}_{\boldsymbol{W}}[\langle \nabla_{y_l} G_l^j(y_l), z_l \rangle^2] = \sum_{k,k'} \mathbb{E}_{\boldsymbol{W}} \left[\frac{\partial G_l^j}{\partial y_l^k} \frac{\partial G_l^j}{\partial y_l^{k'}} \right] \mathbb{E}_{\boldsymbol{W}}[z_l^k z_l^{k'}]$$

Since $\mathbb{E}_{\boldsymbol{W}}[z_l^k z_l^{k'}] = 0$ for $k \neq k'$, we have that

$$\mathbb{E}_{\boldsymbol{W}}[\langle \nabla_{y_l} G_l^j(y_l), z_l \rangle^2] = \sum_{k=1}^{\infty} \mathbb{E}_{\boldsymbol{W}} \left[\left(\frac{\partial G_l^j}{\partial y_l^k} \right)^2 \right] \mathbb{E}_{\boldsymbol{W}}[(z_l^k)^2]$$

Let us deal first with the term $\mathbb{E}_{\boldsymbol{W}}[(z_l^k)^2]$.

Let $k \in \mathbb{N}$. Recall that for finite N, we have that $z_l^k = \sum_{m=1}^{N} W_l^{k,m} \phi(y_{l-1}^m)$ and $W_l^{k,m} \sim \mathcal{N}(0, \frac{2}{N})$. Thus, using the Central Limit Theorem in the limit $N \to \infty$, we obtain

$$\mathbb{E}_{\boldsymbol{W}}[(z_l^k)^2] = 2\, \mathbb{E}_{Z \sim \mathcal{N}(0,1)}[\phi(\sqrt{q_{l-1}} Z)^2] = Q_{l-1}(x, x),$$

where $Q_{l-1}(x, x)$ is given by Lemma A1. We obtain

$$\mathbb{E}_{\boldsymbol{W}}[(z_l^k)^2] = \left(1 + \frac{1}{L}\right)^{l-1} Q_0(x, x).$$

The term $\bar{q}_l^j = \mathbb{E}_{\boldsymbol{W}} \left[\left(\frac{\partial G_l^j}{\partial y_l^k} \right)^2 \right]$ can be computed in a similar fashion to that of the proof of Lemma A3. Indeed, using the same techniques, we obtain

$$\bar{q}_l^{j,k} = \left(1 + \frac{1}{L}\right) \bar{q}_{l+1}^{j,k}$$

which yields

$$\bar{q}_l^{j,k} = \left(1 + \frac{1}{L}\right)^{L-l-1} \bar{q}_L^{j,k}$$

Using the Central Limit Theorem again in the last layer, we obtain

$$\sum_{k=1}^{\infty} \mathbb{E}_{\boldsymbol{W}} \left[\left(\frac{\partial G_l^j}{\partial y_l^k} \right)^2 \right] = \left(1 + \frac{1}{L}\right)^{L-l}.$$

This yields

$$\mathbb{E}_{\boldsymbol{W}}[\langle \nabla_{y_l} G_l^j(y_l), z_l \rangle^2] = \left(1 + \frac{1}{L}\right)^{L-1} Q_0(x, x).$$

and we conclude that

$$\mathbb{E}_{\boldsymbol{W}}[\|\zeta_l(x_i, \boldsymbol{W})\|_2^2] = o \times \left(1 + \frac{1}{L}\right)^{L-1} Q_0(x, x)$$

The latter is independent of l, which concludes the proof.

□

A1.5 Proof of Theorem 2

Consider an arbitrary neuron $y_{\alpha L}^i$ in the $(\alpha L)^{th}$ layer for some fixed $\alpha \in (0, 1)$. $y_{\alpha L}^i(x, \boldsymbol{\delta})$ can be approximated using a first order Taylor expansion around $\boldsymbol{\delta} = \mathbf{1}$. We obtain similarly,

$$y_{\alpha L}^i(x, \boldsymbol{\delta}) \approx \bar{y}_{\alpha L}^i(x) + \frac{1}{\sqrt{L}} \sum_{l=1}^{\alpha L} \eta_l \langle z_l, \nabla_{y_l} G_l^i(y_l(x; \mathbf{1})) \rangle \tag{A10}$$

where G_l^i is defined by $y_{\alpha L}^i(x;\mathbf{1}) = G_l^i(y_l(x;\mathbf{1}))$, $\eta_l = \delta_l - p_l$, and $\bar{y}_{\alpha L}^i(x) = y_{\alpha L}^i(x,\mathbf{1}) + \frac{1}{\sqrt{L}}\sum_{l=1}^{\alpha L}(p_l - 1)\langle z_l, \nabla_{y_l} G_l^i(y_l(x;\mathbf{1}))\rangle \approx y_{\alpha L}^i(x,\boldsymbol{p})$.

Let $\gamma_{\alpha,L}(x) = \frac{1}{\sqrt{L}}\sum_{l=1}^{\alpha L}\eta_l \langle z_l, \nabla_{y_l} G_l^i(y_l(x;\mathbf{1}))\rangle$. The term $\gamma_{\alpha,L}$ captures the randomness of the binary mask $\boldsymbol{\delta}$, which up to a factor α, resembles to the scaled mean in Central Limit Theorem(CLT) and can be written as

$$\gamma_{\alpha,L}(x) = \sqrt{\alpha} \times \frac{1}{\sqrt{\alpha L}} \sum_{l=2}^{\alpha L} X_{l,L}(x)$$

where $X_{l,L}(x) = \eta_l \langle z_l, \nabla_{y_l} G_l^i(y_l(x;\mathbf{1}))\rangle$. We use a Lindeberg's CLT to prove the following result.

Theorem 4. *Let $x \in \mathbb{R}^d$, $X_{l,L}(x) = \eta_l\, \mu_{l,L}(x)$ where $\mu_{l,L}(x) = \langle z_l, \nabla_{y_l} G_l^i(y_l(x;\mathbf{1})\rangle$, and $\sigma_{l,L}^2(x) = Var_\delta[X_{l,L}(x)] = p_l(1-p_l)\mu_{l,L}(x)^2$ for $l \in [L]$. Assume that*

1. *There exists $a \in (0, 1/2)$ such that for all L, and $l \in [L]$, $p_l \in (a, 1-a)$.*
2. $\lim_{L\to\infty} \frac{\max_{k\in[L]} \mu_{k,L}^2(x)}{\sum_{l=1}^L \mu_{l,L}^2(x)} = 0.$
3. $v_{\alpha,\infty}(x) := \lim_{L\to\infty} \frac{\sum_{l=1}^L \sigma_{l,L}^2(x)}{L}$ *exists and is finite.*

Then,

$$\gamma_{\alpha,L}(x) \xrightarrow[L\to\infty]{D} \mathcal{N}(0, \alpha\, v_{\alpha,\infty}(x)).$$

Before proving Theorem 2, we state Lindeberg's CLT for traingular arrays.

Theorem 3 (Lindeberg's Central Limit Theorem for Triangular arrays)**.** *Let $(X_{n,1}, \ldots, X_{n,n})_{n\geq 1}$ be a triangular array of independent random variables, each with finite mean $\mu_{n,i}$ and finite variance $\sigma_{n,i}^2$. Define $s_n^2 = \sum_{i=1}^n \sigma_{n,i}^2$. Assume that for all $\epsilon > 0$, we have that*

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^n \mathbb{E}[(X_{n,i} - \mu_{n,i})^2 1_{\{|X_{n,i}-\mu_{n,i}|>\epsilon s_n\}}] = 0,$$

Then, we have

$$\frac{1}{s_n}\sum_{i=1}^n (X_{n,i} - \mu_{n,i}) \xrightarrow[n\to\infty]{D} \mathcal{N}(0,1)$$

Given an input $x \in \mathbb{R}^d$, the next lemma provides sufficient conditions for the Lindeberg's condition to hold for the triangular array of random variables $X_{l,L}(x) = \eta_l \mu_{l,L}(x)$. In this context, a scaled version of $\gamma_L(x)$ converges to a standard normal variable in the limit of infinite depth.

Lemma A4. *Let $x \in \mathbb{R}^d$, and define $X_{l,L}(x) = \eta_l\, \mu_{l,L}(x)$ where $\mu_{l,L}$ is a deterministic mapping from $\mathbb{R}^d$ to $\mathbb{R}$, and let $\sigma_{l,L}^2(x) = Var_\delta[X_{l,L}(x)]$ for $l \in [L]$. Assume that*

1. *There exists $a \in (0, 1/2)$ such that for all L, and $l \in [L]$, $p_l \in (a, 1-a)$.*
2. $\lim_{L\to\infty} \frac{\max_{k\in[L]} \mu_{k,L}^2(x)}{\sum_{l=1}^L \mu_{l,L}^2(x)} = 0.$

Then for all $\epsilon > 0$, we have that

$$\lim_{L\to\infty} \frac{1}{s_L^2(x)} \sum_{l=1}^L \mathbb{E}[X_{l,L}(x)^2 1_{\{|X_{l,L}(x)|>\epsilon s_L(x)\}}] = 0,$$

where $s_L^2(x) = \sum_{l=1}^L \sigma_{l,L}^2(x)$.

The proof of Lemma A4 is provide in Section A1.6.

Corollary A1. *Under the same assumptions of Lemma A4, we have that*

$$\frac{1}{s_L(x)} \sum_{l=1}^L X_{l,L} \xrightarrow[L\to\infty]{D} \mathcal{N}(0,1)$$

The proof of Theorem 2 follows from Corollary A1.

A1.6 Proof of Lemma A4

Lemma A4. *Let $x \in \mathbb{R}^d$, and define $X_{l,L}(x) = \eta_l \, \mu_{l,L}(x)$ where $\mu_{l,L}$ is a deterministic mapping from $\mathbb{R}^d$ to $\mathbb{R}$, and let $\sigma^2_{l,L}(x) = Var_\delta[X_{l,L}(x)]$ for $l \in [L]$. Assume that*

1. *There exists $a \in (0, 1/2)$ such that for all L, and $l \in [L]$, $p_l \in (a, 1-a)$.*
2. $\lim_{L\to\infty} \frac{\max_{k\in[L]} \mu^2_{k,L}(x)}{\sum_{l=1}^L \mu^2_{l,L}(x)} = 0$.

Then for all $\epsilon > 0$, we have that

$$\lim_{L\to\infty} \frac{1}{s_L^2(x)} \sum_{l=1}^{L} \mathbb{E}[X_{l,L}(x)^2 1_{\{|X_{l,L}(x)|>\epsilon s_L(x)\}}] = 0,$$

where $s_L^2(x) = \sum_{l=1}^L \sigma^2_{l,L}(x)$.

Proof. Fix $i \in [o]$. For $l \in [L]$, we have that $\sigma^2_{l,L} = p_l(1-p_l)\mu^2_{l,L}$. Therefore,

$$s_L^2 = \sum_{l=1}^{L} p_l(1-p_l)\mu^2_{l,L}$$

Under conditions 1 and 2, it is straightforward that $s_L^2 = \Theta(L)$.

To simplify our notation in the rest of the proof, we fix $x \in \mathbb{R}^d$ and denote by $X_l := X_{l,L}(x)$ and $\mu_l := \mu_{l,L}(x)$. Now let us prove the result. Fix $\epsilon > 0$. We have that

$$\mathbb{E}[(X_l)^2 1_{\{|X_i|>\epsilon s_L\}}] = \mu_l^2 \, \mathbb{E}\left[\eta_l^2 1_{\{|\eta_l|>\frac{\epsilon s_L}{|\mu_l|}\}}\right]$$

Using the fact that $|\eta_l| \leq \max(p_l, 1-p_l)$, we have

$$\begin{aligned}\mathbb{E}\left[\eta_l^2 1_{\{|\eta_l|>\frac{\epsilon s_L}{|\mu_l|}\}}\right] &\leq p_l(1-p_l) 1_{\{\max(p_l,1-p_l)>\frac{\epsilon s_L}{|\mu_l|}\}} \\ &\leq p_l(1-p_l)\zeta_L\end{aligned}$$

where $\zeta_L = 1_{\{1-a>\frac{\epsilon s_L}{\max_{k\in[L]}|\mu_k}\}}$.
Knowing that

$$\frac{s_L}{\max_{k\in[L]}|\mu_k|} \geq a\sqrt{\frac{\sum_{l=1}^L \mu_l^2}{\max_{k\in[L]}\mu_k^2}},$$

where the lower bound diverges to ∞ as L increases by assumption. We conclude that

$$\frac{1}{s_L^2}\sum_{l=1}^{L} \mathbb{E}[(X_l)^2 1_{\{|X_l|>\epsilon s_L\}}] \leq \zeta_L \underset{L\to\infty}{\to} 0.$$

. □

A2 Full derivation of explicit regularization with $\mathcal{SD}$

Consider a dataset $\mathcal{D} = \mathcal{X} \times \mathcal{T}$ consisting of n (input, target) pairs $\{(x_i, t_i)\}_{1\leq i\leq n}$ with $(x_i, t_i) \in \mathbb{R}^d \times \mathbb{R}^o$. Let $\ell : \mathbb{R}^d \times \mathbb{R}^o \to \mathbb{R}$ be a smooth loss function, e.g. quadratic loss, crossentropy loss etc. Define the model loss for a single sample $(x, t) \in \mathcal{D}$ by

$$\mathcal{L}(\boldsymbol{W}, x; \boldsymbol{\delta}) = \ell(y_{out}(x; \boldsymbol{\delta}), t), \quad \mathcal{L}(\boldsymbol{W}, x) = \mathbb{E}_\delta\left[\ell(y_{out}(x; \boldsymbol{\delta}), t)\right],$$

where $\boldsymbol{W} = (W_l)_{0\leq l\leq L}$. With $\mathcal{SD}$, we optimize the average empirical loss given by

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\delta\left[\ell(y_{out}(x_i; \boldsymbol{\delta}), t_i)\right]$$

To isolate the regularization effect of $\mathcal{SD}$ on the loss function, we use a second order approximation of the loss function of the model, this will allow us to marginalize out the mask $\boldsymbol{\delta}$. Let $z_l(x;\boldsymbol{\delta}) = \Psi_l(W_l, y_{l-1}(x;\boldsymbol{\delta}))$ be the activations. For some pair $(x,t) \in \mathcal{D}$, the second order Taylor approximation of $\ell(y_L(x),t)$ around $\boldsymbol{\delta} = \mathbf{1} = (1,\dots,1)$ is given by

$$\begin{aligned}\ell(y_{out}(x;\boldsymbol{\delta}),z) \approx \ell(y_{out}(x;\mathbf{1}),z) &+ \frac{1}{\sqrt{L}}\sum_{l=1}^{L}(\delta_l - 1)\langle z_l(x;\mathbf{1}), \nabla_{y_l}[\ell \circ G_l](y_l(x;\mathbf{1}))\rangle \\ &+ \frac{1}{2L}\sum_{l=1}^{L}(\delta_l-1)^2 z_l(x;\mathbf{1})^T \nabla^2_{y_l}[\ell\circ G_l](y_l(x;\mathbf{1}))z_l(x;\mathbf{1})\end{aligned} \tag{A11}$$

where G_l is the function defined by $y_{out}(x;\mathbf{1}) = G_l(y_{l-1}(x;\mathbf{1}) + \frac{1}{\sqrt{L}}z_l(x;\mathbf{1}))$. Taking the expectation with respect to $\boldsymbol{\delta}$, we obtain

$$\mathcal{L}(\boldsymbol{W},x) \approx \bar{\mathcal{L}}(\boldsymbol{W},x) + \frac{1}{2L}\sum_{l=1}^{L} p_l(1-p_l)g_l(\boldsymbol{W},x) \tag{A12}$$

where $\bar{\mathcal{L}}(\boldsymbol{W},x) \approx \ell(y_{out}(x;\boldsymbol{p}),t)$ (more precisely, $\bar{\mathcal{L}}(\boldsymbol{W},x)$ is the second order Taylor approximation of $\ell(y_{out}(x;\boldsymbol{p}),t)$ around $\boldsymbol{p} = 1$[5]), and $g_l(\boldsymbol{W},x) = z_l(x;\mathbf{1})^T\nabla^2_{y_l}[\ell\circ G_l](y_l(x;\mathbf{1}))z_l(x;\mathbf{1})$.

The first term $\bar{\mathcal{L}}(\boldsymbol{W},x)$ in Eq. (4) is the loss function of the average network (i.e. replacing $\boldsymbol{\delta}$ with its mean $\boldsymbol{p}$). Thus, Eq. (4) shows that training with $\mathcal{SD}$ entails training the average network with an explicit regularization term that implicitly depends on the weights $\boldsymbol{W}$.

A3 Implicit regularization and gradient noise

The results of Theorem 2 can be generalized to the gradient noise. Adding noise to the gradient is a well-known technique to improve generalization. It acts as an implicit regularization on the loss function. Neelakantan et al. [2015] suggested adding a zero mean Gaussian noise parameterized by its variance. At training time t, this translates to replacing $\frac{\partial\mathcal{L}}{\partial w_t}$ by $\frac{\partial\mathcal{L}}{\partial w_t} + \mathcal{N}(0,\sigma_t^2)$, where w_t is the value of some arbitrary weight in the network at training time t, and $\sigma_t^2 = a(1+t)^{-b}$ for some constants $a, b > 0$. As t grows, the noise converge to 0 (in ℓ_2 norm), letting the model stabilize in a local minimum. Empirically, adding this noise tends to boost the performance by making the model robust to over-fitting. Using similar perturbation analysis as in the previous section, we show that when the depth is large, *$\mathcal{SD}$mimics this behaviour by implicitly adding a Gaussian noise to the gradient at each training step.*

Consider an arbitrary weight w in the network, and let $h(x;\boldsymbol{\delta}) = \frac{\partial\mathcal{L}(x,\boldsymbol{W};\boldsymbol{\delta})}{\partial w}$ be the gradient of the model loss w.r.t w. $h(x;\boldsymbol{\delta})$ can be approximated using a first order Taylor expansion of the loss around $\boldsymbol{\delta} = \mathbf{1}$. We obtain,

$$\begin{aligned}h(x;\boldsymbol{\delta}) &= \frac{\partial\mathcal{L}(x,\boldsymbol{W};\boldsymbol{\delta})}{\partial w} \\ &\approx \frac{\partial}{\partial w}\left(\mathcal{L}(x,\boldsymbol{W};\mathbf{1}) + \frac{1}{\sqrt{L}}\sum_{l=1}^{L}(\delta_l-1)\langle z_l, \nabla_{y_l}[\ell\circ G_l](y_l(x;\mathbf{1}))\rangle\right) \\ &\approx \bar{h}(x) + \frac{1}{\sqrt{L}}\sum_{l=1}^{L}\eta_l\frac{\partial}{\partial w}\langle z_l, \nabla_{y_l}[\ell\circ G_l](y_l(x;\mathbf{1}))\rangle\end{aligned} \tag{A13}$$

where $\eta_l = \delta_l - p_l$, and $\bar{h}(x) = h(x;\mathbf{1}) + \frac{1}{\sqrt{L}}\sum_{l=1}^{L}(p_l-1)\frac{\partial}{\partial w}\langle z_l, \nabla_{y_l}[\ell\circ G_l](y_l(x;\mathbf{1}))\rangle \approx h(x;\boldsymbol{p})$.

Let $\gamma_L(x) = \frac{1}{\sqrt{L}}\sum_{l=1}^{L}\eta_l\frac{\partial}{\partial w}\langle z_l, \nabla_{y_l}[\ell\circ G_l](y_l(x;\mathbf{1}))\rangle$. With $\mathcal{SD}$, the gradient $h(x;\boldsymbol{\delta})$ can therefore be seen as a perturbation of the gradient of the average network $h(x;\boldsymbol{p})$ with a noise encoded by

[5]Note that we could obtain Eq. (4) using the Taylor expansion around $\boldsymbol{\delta} = \boldsymbol{p}$. However, in this case, the hessian will depend on $\boldsymbol{p}$, which complicates the analysis of the role of $\boldsymbol{p}$ in the regularization term.

$\gamma_L(x)$. The scaling factor $1/\sqrt{L}$ ensures that γ_L remains bounded (in ℓ_2 norm) as L grows. Without this scaling, the variance of γ_L generally explodes. The term γ_L captures the randomness of the binary mask $\boldsymbol{\delta}$, which resembles to the scaled mean in Central Limit Theorem and can be written as

$$\gamma_L(x) = \frac{1}{\sqrt{L}} \sum_{l=2}^{L} X_{l,L}(x)$$

where $X_{l,L}(x) = \eta_l \frac{\partial}{\partial w} \langle z_l, \nabla_{y_l}[\ell \circ G_l](y_l(x;\mathbf{1}))\rangle$. Ideally, we would like to apply Central Limit Theorem (CLT) to conclude on the Gaussianity of $\gamma_L(x)$ in the large depth limit. However, the random variables X_l are generally not $i.i.d$ (they have different variances) and they also depend on L. Thus, standard CLT argument fails. Fortunately, there is a more general form of CLT known as Lindeberg's CLT which we use in the proof of the next theorem.

Theorem 4 (Asymptotic normality of gradient noise)**.** *Let $x \in \mathbb{R}^d$, and define $X_{l,L}(x) = \eta_l \, \mu_{l,L}(x)$ where $\mu_{l,L}(x) = \frac{\partial}{\partial w} \langle z_l, \nabla_{y_l}[\ell \circ G_l](y_l(x;\mathbf{1})\rangle$, and let $\sigma^2_{l,L}(x) = Var_\delta[X_{l,L}(x)] = p_l(1-p_l)\mu_{l,L}(x)^2$ for $l \in [L]$. Assume that*

1. *There exists $a \in (0, 1/2)$ such that for all L, and $l \in [L]$, $p_l \in (a, 1-a)$.*
2. $\lim_{L\to\infty} \frac{\max_{k\in[L]} \mu^2_{k,L}(x)}{\sum_{l=1}^{L} \mu^2_{l,L}(x)} = 0.$
3. $v_\infty(x) := \lim_{L\to\infty} \frac{\sum_{l=1}^{L} \sigma^2_{l,L}(x)}{L}$ *exists and is finite.*

Then,

$$\gamma_L(x) \xrightarrow[L\to\infty]{D} \mathcal{N}(0, v_\infty(x))$$

As a result, $\mathcal{SD}$ implicitly mimics regularization techniques that adds Gaussian noise to the gradient.

The proof of Theorem 4 follows from Lemma A4 in a similar fashion to the proof of Theorem 2. Under the assumptions of Theorem 4, training a ResNet with $\mathcal{SD}$ entails adding a gradient noise $\gamma_L(x)$ that becomes asymptotically close (in distribution) to a Gaussian random variable. The limiting variance of this noise, given by $v_\infty(x)$, depends on the input x, which arises the question of the nature of the noise process $\gamma_L(.)$. It turns out that under some assumptions, $\gamma_L(.)$ converges to a Gaussian process in the limit of large depth. We show this in the next proposition

A4 Further experimental results

Implementation details: Vanilla Stable ResNet is composed of identical residual blocks each formed of a Linear and a ReLu layer. Stable ResNet110 follows [He et al., 2016, Huang et al., 2016]; it comprises three groups of residual blocks; each block consists of a sequence of layers Convolution-BatchNorm-ReLU-Convolution-BatchNorm. We build on an open-source implementation of standard ResNets[6]. We scale the blocks using a factor $1/\sqrt{L}$ as described in Section 3. We use the adjective non-stable to qualify models where the scaling is not performed. The toy regression task consists of estimating the function

$$f_\beta : x \mapsto \sin(\beta^T x),$$

where the inputs x and parameter β are in $\mathbb{R}^{256}$, sampled from a standard Gaussian. The output is unidimensional. CIFAR-10, CIFAR-100 contain 32-by-32 color images, representing respectively 10 and 100 classes of natural scene objects. The models are learned in 164 epochs. The Stable ResNet56 and ResNet110 use an initial learning rate of 0.01, divided by 10 at epochs 80 and 120. Parameter optimization is conducted with SGD with a momentum of 0.9 and a batch size of 128. The Vanilla ResNet models have an initial learning rate of 0.05, and a batch size of 256. We use 4 GPUs V100 to conduct the experiments. In the results of Section 3 and 4, the expectations are empirically evaluated using 500 Monte-Carlo (MC) samples. The boxplots are also obtained using 500 MC samples.

The exploding gradient of Non-Stable Vanilla ResNet: In Table 6 and Table 5, we empirically validate Proposition 1. We compare the empirical values of

$$\frac{1}{L-l}\log \tilde{q}_l(x,z) = \frac{1}{L-l}\log \mathbb{E}_{W,\boldsymbol{\delta}} \frac{\|\nabla_{y_l}\mathcal{L}\|^2}{\|\nabla_{y_L}\mathcal{L}\|^2},$$

and compare it to the theoretical value (in parenthesis in the tables). We consider two different survival proportions. We see an excellent match between the theoretical value and the empirical one.

Proposition 1 coupled to the concavity of $\log(1+x)$ implies that at a constant budget, the uniform rate is the mode that suffers the most from gradient explosion. Figures 2a and 2b illustrate this phenomenon. We can see that the gradient magnitude of the uniform mode can be two orders of magnitude larger than in the linear case. However, the Stable scaling alleviates this effect; In Figure 2c we can see that none of the modes suffers from the gradient explosion anymore.

Second order approximation of the loss: In Table 3 we empirically verify the approximation accuracy of the loss (equation (6)). $\mathcal{L}$ is the loss that is minimized when learning with $\mathcal{SD}$. $\bar{\mathcal{L}}$ is the loss of the average model $y_{out}(x;\boldsymbol{p})$. The penalization term is $\frac{1}{2L}\sum_{l=1}^{L} p_l(1-p_l)g_l(\boldsymbol{W})$ (more details in Section 4). At initialization, the loss of the average model accurately represents the SD loss; the penalization term only brings a marginal correction. As the training goes, the penalization term becomes crucial; $\bar{\mathcal{L}}$ only represents 12% of the loss after convergence. We can interpret this in the light of the fact that $\bar{\mathcal{L}}$ converges to zero, whereas the penalization term does not necessarily do. We note that the second-order approximation does not capture up to 25% of the loss. We believe that this is partly related to the non-PSD term Γ_l that we discarded for the analysis.

Table 3: Empirical verification of Equation (7) with Vanilla Resnet50 with width 256 and average survival probability $\bar{L}/L = 0.8$ with uniform mode.

epoch	$\left\lvert\frac{\mathcal{L}-\bar{\mathcal{L}}}{\mathcal{L}}\right\rvert$	$\left\lvert\frac{\mathcal{L}-\bar{\mathcal{L}}-pen}{\mathcal{L}}\right\rvert$	Ratio
0	0.015	0.003	$\times 5.7$
40	0.389	0.084	$\times 4.6$
80	0.651	0.183	$\times 3.5$
120	0.856	0.231	$\times 3.7$
160	0.884	0.245	$\times 3.6$

Further empirical verification of assumption 2 of Theorem 2: Under some assumptions, Theorem 2 guarantees the asymptotic normality of the noise γ. Further empirical verifications of assumption 2 are shown in Fig. 6 and Fig. 7. The downtrend is consistent throughout training and modes, suggesting that assumption 2 is realistic. In Fig. 8 we plot the distributions of the pvalues of two normality tests: the Shapiro–Wilk (Shapiro and Wilk [1965]) test and the the D'Agostino's K^2-tests (D'Agostino [1970]).

[6] https://github.com/felixgwu/img_classification_pk_pytorch

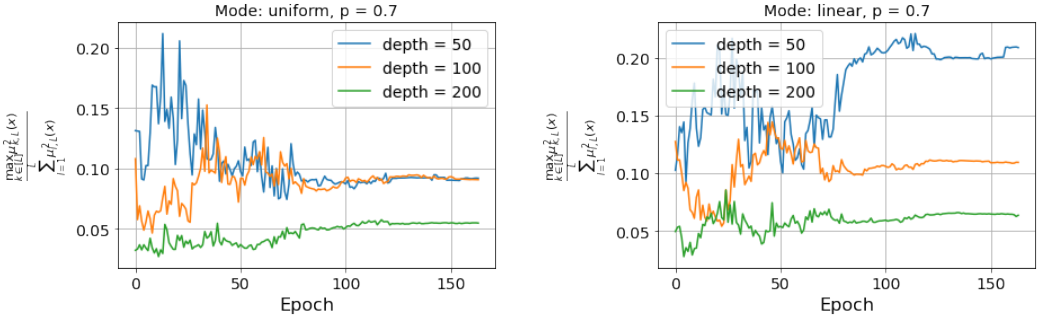
Further empirical verification of the Budget Hypothesis: We also compare the three modes (Uniform, Linear and SenseMode) in CIFAR10 and CIFAR100 for Stable Resnet56. The results are reported in Table 6. These results confirm the observations discussed in the main text.

Table 4: Empirical verification of Proposition 1 with Vanilla Resnet50 with width 512 and average survival probability $\bar{L}/L = 0.5$. Comparison between the empirical average growth rate of the gradient magnitude against the theoretical value (between parenthesis) at initialization.

ℓ	Standard	Uniform	Linear
0	2.003 (2)	1.507 (1.5)	1.433 (1.473)
10	2.002 (2)	1.499 (1.5)	1.349 (1.374)
20	2.001 (2)	1.502 (1.5)	1.248 (1.284)
30	2.002 (2)	1.504 (1.5)	1.207 (1.191)
40	2.002 (2)	1.542 (1.5)	1.079 (1.097)

Table 5: Empirical verification of Proposition 1 with Vanilla Resnet50 with width 512 and average survival probability $\bar{L}/L = 0.7$. Comparison between the empirical average growth rate of the gradient magnitude against the theoretical value (between parenthesis) at initialization.

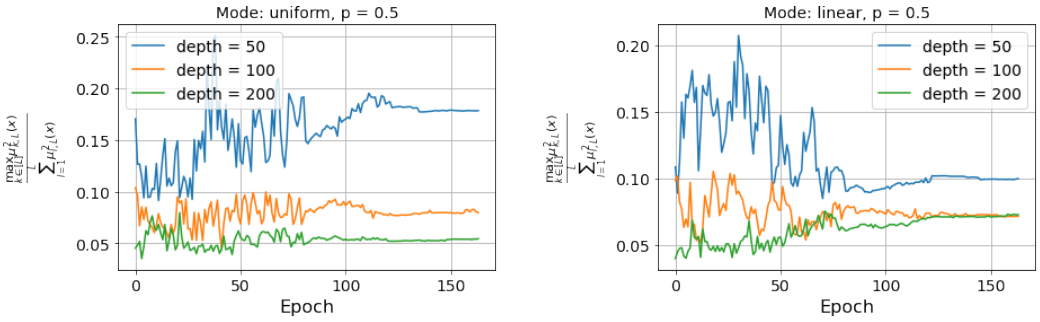
ℓ	Standard	Uniform	Linear
0	2.001 (2)	1.705 (1.7)	1.694 (1.691)
10	2.001 (2)	1.708 (1.7)	1.633 (1.629)
20	2.001 (2)	1.707 (1.7)	1.569 (1.573)
30	2.001 (2)	1.716 (1.7)	1.555 (1.516)
40	1.999 (2)	1.739 (1.7)	1.530 (1.459)



(a) Uniform mode

(b) Linear mode

Figure 6: Empirical verification of assumption 2 of Theorem 2 on Vanilla ResNet with width 256 with average survival probability $\bar{L}/L = 0.7$.



(a) Uniform mode

(b) Linear mode

Figure 7: Empirical verification of assumption 2 of Theorem 2 on Vanilla ResNet with width 256 with average survival probability $\bar{L}/L = 0.5$.

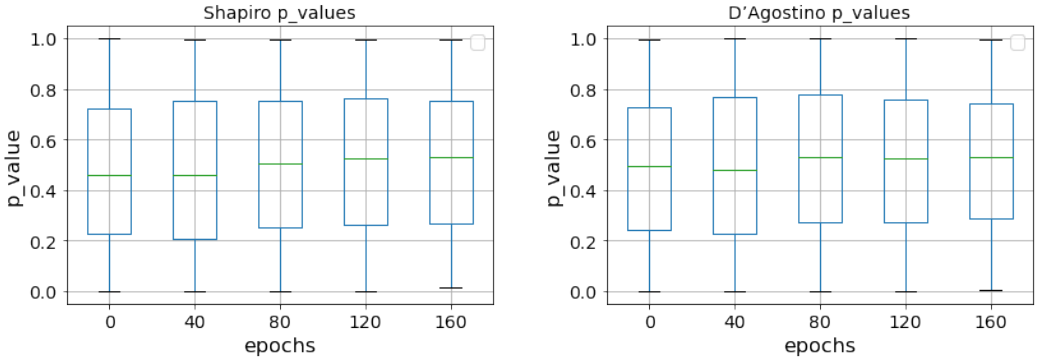


Figure 8: Empirical verification of Theorem 2 on Vanilla ResNet100 with width 128 with average survival probability $\bar{L}/L = 0.7$ and uniform mode. Distribution of the p-values for two normality tests: Shapiro and D'Agostino's tests

Table 6: Comparison of the modes of selection of the survival probabilities with fixed budget with Stable ResNet56.

$\bar{L}/L$	Uniform	SenseMode	Linear
0.1	24.58 ± 0.3	**2.93** ± 0.4	–
0.2	13.85 ± 0.3	**11.72** ± 0.3	–
0.3	10.23 ± 0.2	**8.59** ± 0.4	–
0.4	**8.49** ± 0.2	**8.23** ± 0.3	–
0.5	**8.38** ± 0.2	**8.25** ± 0.3	12.01 ± 0.3
0.6	**7.34** ± 0.3	8.17 ± 0.2	9.26 ± 0.2
0.7	**8.03** ± 0.1	8.20 ± 0.1	8.30 ± 0.1
0.8	**6.48** ± 0.1	7.55 ± 0.1	6.89 ± 0.2
0.9	7.16 ± 0.1	7.81 ± 0.1	**6.62** ± 0.1
1		7.10 ± 0.1	

(a) CIFAR10 with ResNet56

$\bar{L}/L$	Uniform	SenseMode	Linear
0.1	61.98 ± 0.3	$60,27 \pm 0.2$	–
0.2	47.24 ± 0.2	45.74 ± 0.3	–
0.3	39.38 ± 0.2	$37,11 \pm 0.2$	–
0.4	$35,54 \pm 0.2$	$33,71 \pm 0.4$	–
0.5	32.32 ± 0.1	31 ± 0.3	$40,71 \pm 0.2$
0.6	29.57 ± 0.1	30.19 ± 0.3	34.13 ± 0.1
0.7	28.49 ± 0.4	29.69 ± 0.1	30.14 ± 0.1
0.8	27.23 ± 0.2	29.31 ± 0.2	28.34 ± 0.2
0.9	27.01 ± 0.1	$29,45 \pm 0.2$	$27,35 \pm 0.2$
1		28.93 ± 0.5	

(b) Cifar100 with ResNet56