Stable ResNet

S. Hayou, E. Clerico, B. He,
G. Deligiannidis, A. Doucet, J. Rousseau

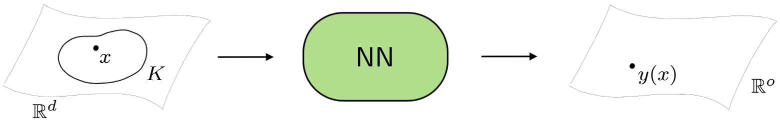
University of Oxford, Department of Statistics

AISTATS 2021

Residual Architectures

- Residual connections map the identity between layers
- Information can easily propagate through the network
- Allow for deeper architectures with improved performance
- Most of sota neural networks have residual connections
- Allow for convolutional layers, batch normalization...

Vanilla ReLU ResNet



$$y_0(x) = \frac{\sigma_w}{\sqrt{d}} W_0\, x + \sigma_b\, B_0$$
$$y_l(x) = y_{l-1}(x) + \frac{\sigma_w}{\sqrt{N_{l-1}}} W_l \phi(y_{l-1}(x)) + \sigma_b B_{l+1}$$
$$y(x) = F(y_L(x))$$

$$y_l \in \mathbb{R}^{N_l}$$
$$W_l \in \mathbb{R}^{N_{l-1}\times N_l}$$
$$B_l \in \mathbb{R}^{N_l}$$
$$\phi(y) = \max(0, y)$$

Random initialization: the components of the parameters are iid.

$$W_l^{ij} \sim \mathcal{N}(0,1) \qquad\qquad B_l^i \sim \mathcal{N}(0,1)$$

Gaussian Limit

- Limit of infinite width: for each layer $N_l \to \infty$
- Each node of the l-th layer is a sum of N_{l-1} iid RVs reweighted by $1/\sqrt{N_{l-1}}$
  $\implies$ Each node is a Gaussian RV
- Each node depends on the input $\implies$ Gaussian Process (GP)

$$y_l^i(\cdot) \sim Y_l(\cdot) \sim \mathcal{GP}(0, Q_l)$$

Gaussian Limit

- What is a Gaussian Process?

  $(Y(x))_{x\in\mathcal{X}} \sim \mathcal{GP}(0, Q)$:
  - $(Y(x_1)\dots Y(x_n))$ jointly normally distributed
  - $\mathbb{E}[Y(x)] = 0$
  - $\mathbb{E}[Y(x)Y(x')] = Q(x, x')$

- How to determine the kernels Q_l? Recursion...

$$Q_0(x, x') = \frac{\sigma_w^2}{d}\, x \cdot x' + \sigma_b^2$$

$$Q_{l+1}(x, x') = \sigma_w^2 \underbrace{\mathbb{E}[\phi(Y_l(x))\phi(Y_l(x'))]}_{\mathcal{F}(Q_l(x,x'),\, C_l(x,x'))} + \sigma_b^2$$

$$C_l(x, x') = Q_l(x, x') / \sqrt{Q_l(x, x) Q_l(x', x')}$$

Infinite-Depth Limit $L \to \infty$

- Kernel (and gradient) explosion:

  - The covariances explode with the depth

  $$Q_l(x,x) \geq \left(1 + \frac{\sigma_w^2}{2}\right)^l \left(\sigma_b^2 \left(1 + \frac{2}{\sigma_w^2}\right) + \frac{\sigma_w^2}{d}\|x\|^2\right)$$

  - The gradient of the loss explodes with depth so the net cannot be trained.

- Inexpressivity:

  The correlation C_l becomes constant for large l:

  $$C_l(x,x') \to 1 \quad \forall x, x'$$

  The output is trivial since the correlation has no dependence on the input.

How to fix these issues?

- At initialization each layer is adding an independent random noise
- All these noisy contributions sum up and bring about the divergence

GOAL: Control the noisy contribution of each layer
SOLUTION: Introduction of scaling factors

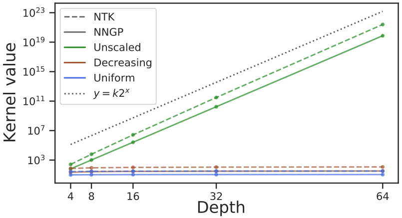
Adding scaling factors $\{\lambda_{l,L}\}_{l\in 1:L}$

$$y_0(x) = \frac{\sigma_w}{\sqrt{d}} W_0 x + \sigma_b B_0$$
$$y_l(x) = y_{l-1}(x) + \lambda_{l,L} \times \left(\frac{\sigma_w}{\sqrt{N_{l-1}}} W_l \phi(y_{l-1}(x)) + \sigma_b B_{l+1}\right)$$
$$y(x) = F(y_L(x))$$

Proposition

$\lim_{L\to\infty} \sum_{l=1}^{L} \lambda_{l,L}^2 < \infty \iff$ Stability of the infinite-depth limit

2 simple cases:
- Uniform Scaling: $\lambda_{l,L} = 1/\sqrt{L}$
- Decreasing Scaling: $\lambda_{l,L} = \lambda_l$ decreasing square-summable sequence



Expressivity

Full expressivity on a compact K (2 equivalent definitions):

- The network can approximate any function in $L^2(K)$ with arbitrary precision and non-zero probability
- The output covariance kernel is universal (RKHS is dense in $C(K)$)

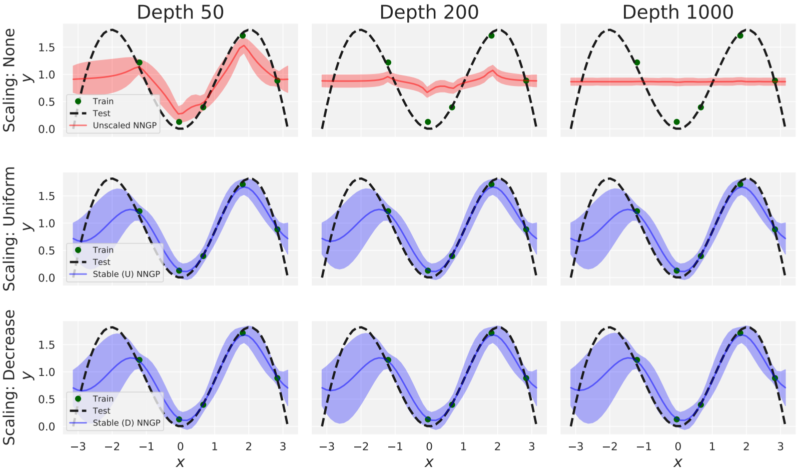
Theorem

Stable ResNets are fully expressive on any compact if $\sigma_b > 0$.
Stable ResNets are fully expressive on the sphere if $\sigma_b = 0$.

Benefits of the scaling

STABLE RESNET $(L = \infty)$	UNSCALED RESNET $(L = \infty)$
Stable at init	Exploding at init
Bounded grad at init	Exploding grad at init
Fully expressive NNGP	Trivial NNGP
Fully expressive NTK	Trivial NTK
Trainable	Untrainable

Experiments: NNGP Regression



Experiments: Image Classification

Dataset	Depth	Scaled (D)	Scaled (U)	Unscaled
C-10	32	$94.84_{\pm 0.08}$	$94.78_{\pm 0.17}$	$94.66_{\pm 0.07}$
	50	$\mathbf{95.07}_{\pm 0.06}$	$94.99_{\pm 0.03}$	$94.85_{\pm 0.06}$
	104	$95.14_{\pm 0.19}$	$\mathbf{95.31}_{\pm 0.07}$	$95.10_{\pm 0.21}$
C-100	32	$\mathbf{75.06}_{\pm 0.05}$	$74.79_{\pm 0.28}$	$74.01_{\pm 0.14}$
	50	$\mathbf{76.20}_{\pm 0.22}$	$75.81_{\pm 0.20}$	$74.66_{\pm 0.33}$
	104	$\mathbf{77.44}_{\pm 0.09}$	$76.88_{\pm 0.39}$	$75.08_{\pm 0.42}$
Tiny-I	32	$63.01_{\pm 0.22}$	$\mathbf{63.06}_{\pm 0.04}$	$62.79_{\pm 0.08}$
	50	$64.78_{\pm 0.24}$	$64.74_{\pm 0.10}$	$63.96_{\pm 0.39}$
	104	$66.57_{\pm 0.39}$	$66.67_{\pm 0.12}$	$65.27_{\pm 0.52}$

Final remarks

- Similar results hold for the NTK. The NTK of a Standard ResNet explode with depth and becomes trivial. Conversely for a Stable ResNet it is bounded and fully expressive.
- We derived a PAC-Bayesian bound for NNGP regression which diverges with depth for a Standard ResNet and keeps bounded for a Stable ResNet.
- In our experiments deep ResNets with either the decreasing or the uniform scaling outperforms the standard architecture. However the selection of an optimal scaling remains an open question.
- Future work directions: links with batch normalization, general activation function, optimal scaling...