Path-conditioned training: a principled way to rescale ReLU neural networks

Arthur Lebeurrier, Titouan Vayer, Rémi Gribonval

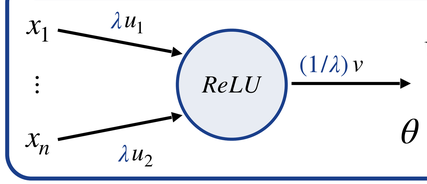
ICML International Conference On Machine Learning

arXiv

Université de Lyon · ENS de Lyon · Inria

TL; DR: We exploit the rescaling symmetry at initialization to accelerate optimization

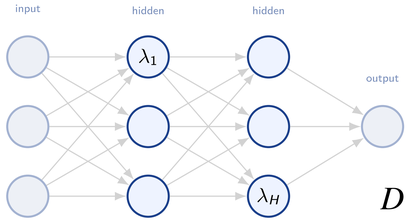
Symmetry in a ReLU Neuron: **Rescaling**



$$v\rho\Big(\sum_i u_i x_i\Big) = \frac{1}{\lambda} v\rho\Big(\sum_i \lambda u_i x_i\Big)$$

$$\theta \mapsto D\theta \qquad D = D_{\lambda,v} \in \mathscr{D}$$

Rescale an entire ReLU Network



The rescaling conserves the network realization

$$f_\theta = f_{D\theta}$$

$$D = D_{\lambda_1,v_1} \times \cdots \times D_{\lambda_H,v_H} \in \mathscr{D}$$

Rescaling matrices $\mathscr{D}$

$$\mathscr{D} = \{\exp(Bu), B \in \{0,1\}^{p\times H}, u \in \mathbb{R}^H\}$$

Characterisation:

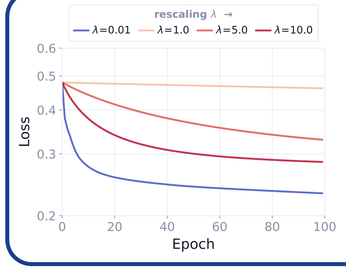
$$D \Leftrightarrow \lambda_1, \cdots, \lambda_H$$

Equivalence class $[\theta]$

$$\theta \sim \theta' \Leftrightarrow \exists D \in \mathscr{D}, \theta' = D\theta$$

Property:

$$\theta \sim \theta' \Rightarrow f_\theta = f_{\theta'}$$

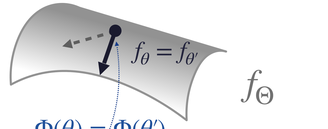


Gradient Flow (GF) $\dot\theta = -\nabla L(\theta)$

$\theta_0' \sim \theta_0$ at initialization

does not imply the same loss decay

Hypothesis: lifted space is a good proxy for functional space

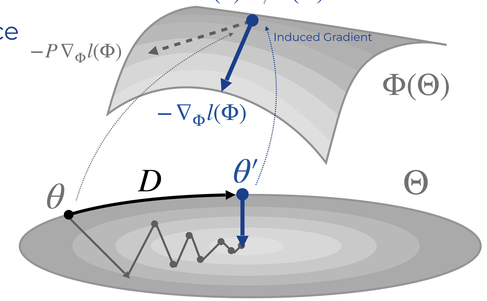


Induced ODE in lifted space

$$\partial_t \Phi(\theta) = -\underbrace{P_\theta}_{\partial\Phi\partial\Phi^\top} \nabla_\Phi \ell(\Phi(\theta))$$

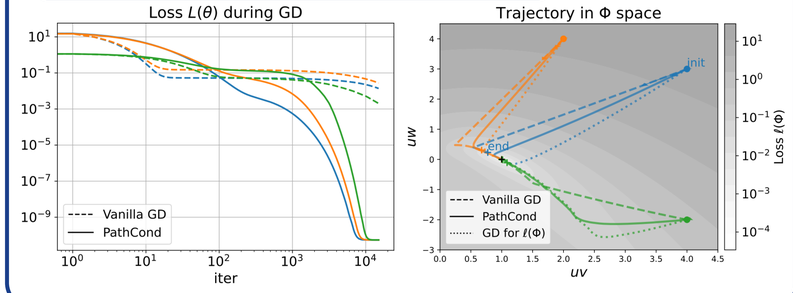
GF in parameter space

$$\dot\theta = -\nabla L(\theta)$$



Contribution: the first *principled* heuristic to choose the *best representative* in $[\theta]$

Mimic Gradient Flow in the lifted space



Reformulation

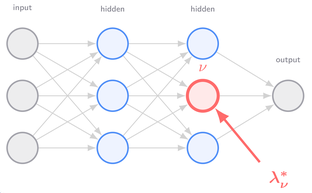
$$P_{\theta'}^{1} \approx I \quad \Leftrightarrow \quad \min_D d(P_{\theta'}\|I)$$

$$\Leftrightarrow \min_{u_h} F(u_1, \ldots, u_h, \cdots, u_H)$$

Close form solution

F **Convex function**

PathCond Algorithm

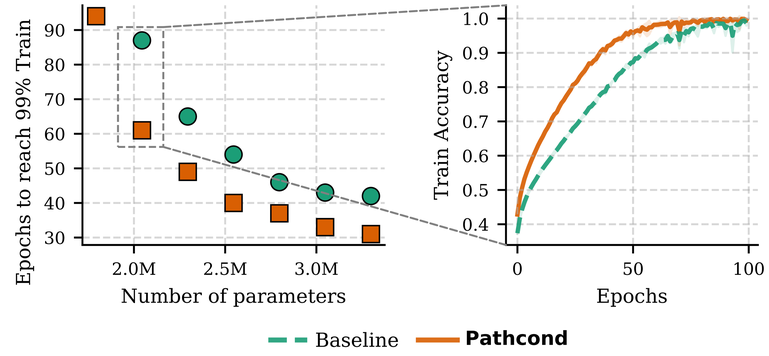


BCD[2] on hidden neurons at cost $O(\#\text{params})$

[1] $P_{\theta'} = \partial\Phi(\theta')\partial\Phi(\theta')^\top$

[2] Block Coordinate Descent

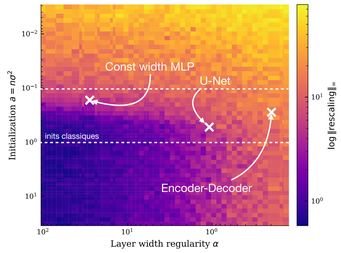
Experiments



Favorable Regimes

1. Non constant layer width

2. Small variance init



Key Take-Aways

1. GF **is not** rescaling invariant
2. **PathCond** is the first principled heuristic for choosing a rescaling
3. **Code available** on github.com/artim436

[1] Stock, Pierre, et al. "Equi-normalization of neural networks." ICLR (2019). [2] Zhao, Bo, et al. "Symmetry teleportation for accelerated optimization." NeurIPS (2022) [3] Lebeurrier, Arthur, Vayer Titouan, and Gribonval Rémi. "Path-conditioned training: a principled way to rescale ReLU neural networks." ICML (2026).