How degenerate is the parametrization of neural networks with the ReLU activation function?

Julius Berner, Dennis Elbrächter, Philipp Grohs

University of Vienna

Notation

- set $\mathcal{P}_N$ of *parametrizations* with architecture $N := (d, N_1, \dots, N_{L-1}, D) \in \mathbb{N}^{L+1}$

$$\mathcal{P}_N := \prod_{\ell=1}^{L} \left(\mathbb{R}^{N_\ell \times N_{\ell-1}} \times \mathbb{R}^{N_\ell}\right)$$

- affine mappings $W_\ell(x) := A_\ell x + b_\ell$ and *realization* map

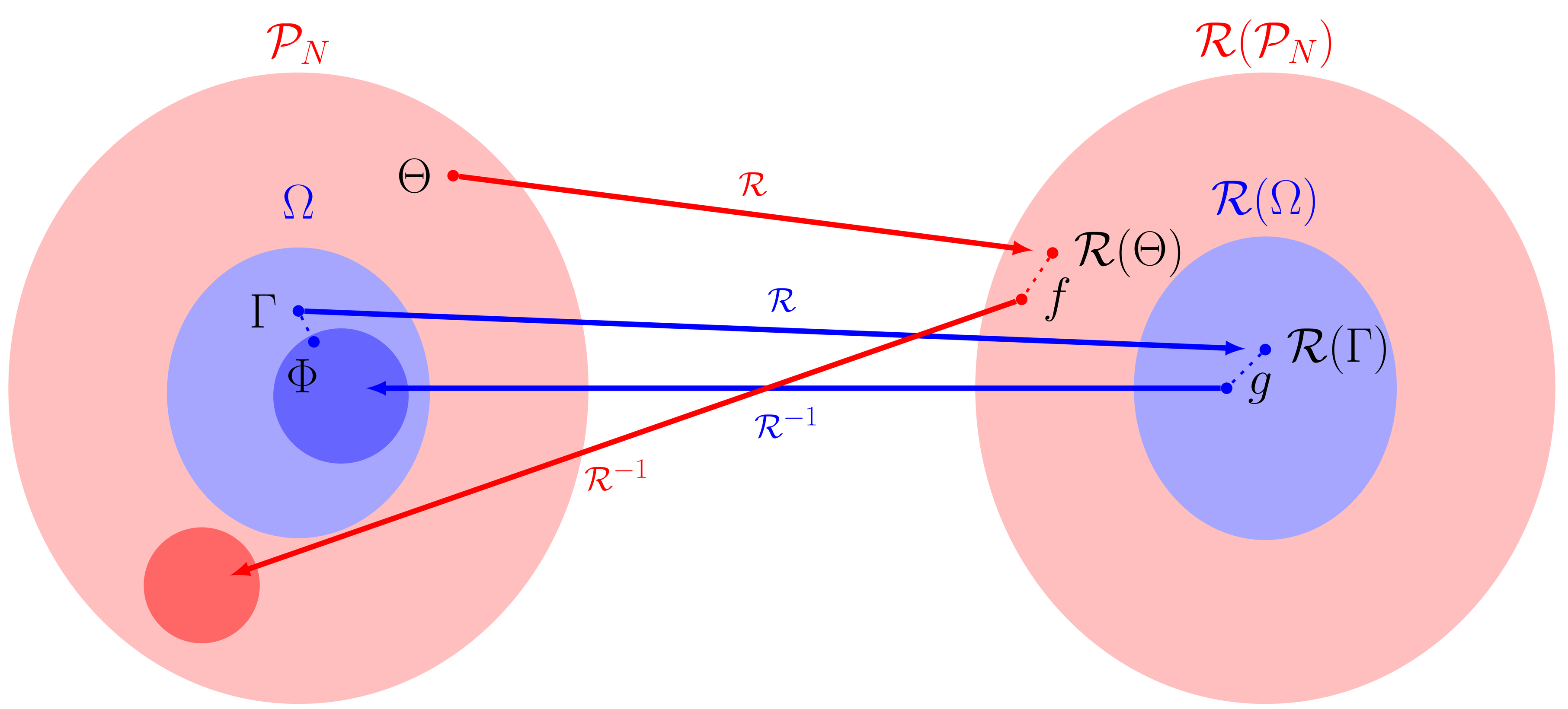
$$\mathcal{R}\colon \mathcal{P}_N \to C(\mathbb{R}^d, \mathbb{R}^D)$$
$$\Theta = ((A_\ell, b_\ell))_{\ell=1}^{L} \mapsto \mathcal{R}(\Theta) := W_L \circ \mathrm{ReLU} \circ W_{L-1} \dots \mathrm{ReLU} \circ W_1$$

Inverse Stability

"Given $\mathcal{R}(\Gamma)$ and $\mathcal{R}(\Theta)$ that are close, does there exist a parametrization Φ with $\mathcal{R}(\Phi) = \mathcal{R}(\Theta)$ such that Γ and Φ are close?"

Definition. Let $\Omega \subseteq \mathcal{P}_N$. The realization map is (s, α) *inverse stable* on Ω w.r.t. $\|\cdot\|$, if for every $\Gamma \in \Omega$ and $g \in \mathcal{R}(\Omega)$ there exists $\Phi \in \Omega$ with

$$\mathcal{R}(\Phi) = g \quad \text{and} \quad \|\Phi - \Gamma\|_\infty \le s\|g - \mathcal{R}(\Gamma)\|^\alpha.$$



References

[1] Julius Berner, Dennis Elbrächter, and Philipp Grohs. "How degenerate is the parametrization of neural networks with the ReLU activation function?" In: *arXiv:1905.09803* (2019).

[2] Julius Berner et al. "Towards a regularity theory for ReLU networks – chain rule and global error estimates". In: *arXiv:1905.04992* (2019). Accepted for presentation at SampTA 2019.

Implications of Inverse Stability

"Local minima in the parameter space induce local minima in the realization space."

Implication 1. Let $\mathcal{L}\colon C(\mathbb{R}^d, \mathbb{R}^D) \to \mathbb{R}_+$ be a *loss function* and $\Gamma_* \in \Omega$ be a local min. of $\mathcal{L} \circ \mathcal{R}$ on Ω with radius $r > 0$, then $\mathcal{R}(\Gamma_*)$ is a local min. of $\mathcal{L}$ on $\mathcal{R}(\Omega)$ with radius $r' = (\frac{r}{s})^{1/\alpha}$.

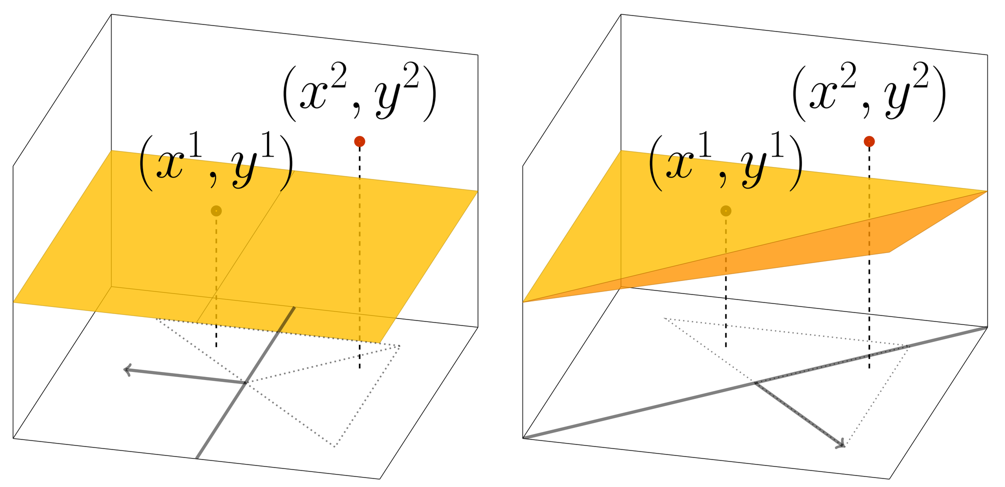


Figure: counter-example without inverse stability

Main Theorem

The realization map is $(4, \frac{1}{2})$ inverse stable w.r.t. $|\cdot|_{W^{1,\infty}}$ on $\mathcal{N}_N := \{\Theta \in \mathcal{P}_N \colon \Theta = \left(([a_1|\dots|a_m]^T, 0), ([c_1|\dots|c_m], 0)\right)\}$ (biasless two-layer networks) regularized such that

- the weights are balanced, i.e. $\|a_i\|_\infty = \|c_i\|_\infty$,
- there are no redundant directions, i.e. $a_i \not\parallel a_j$, and
- the last two coordinates of each a_i are strictly positive.

"For sufficiently large architectures the local minima of a regularized neural network optimization problem are almost optimal."

Implication 2. Let Λ be a quasi-convex regularizer and

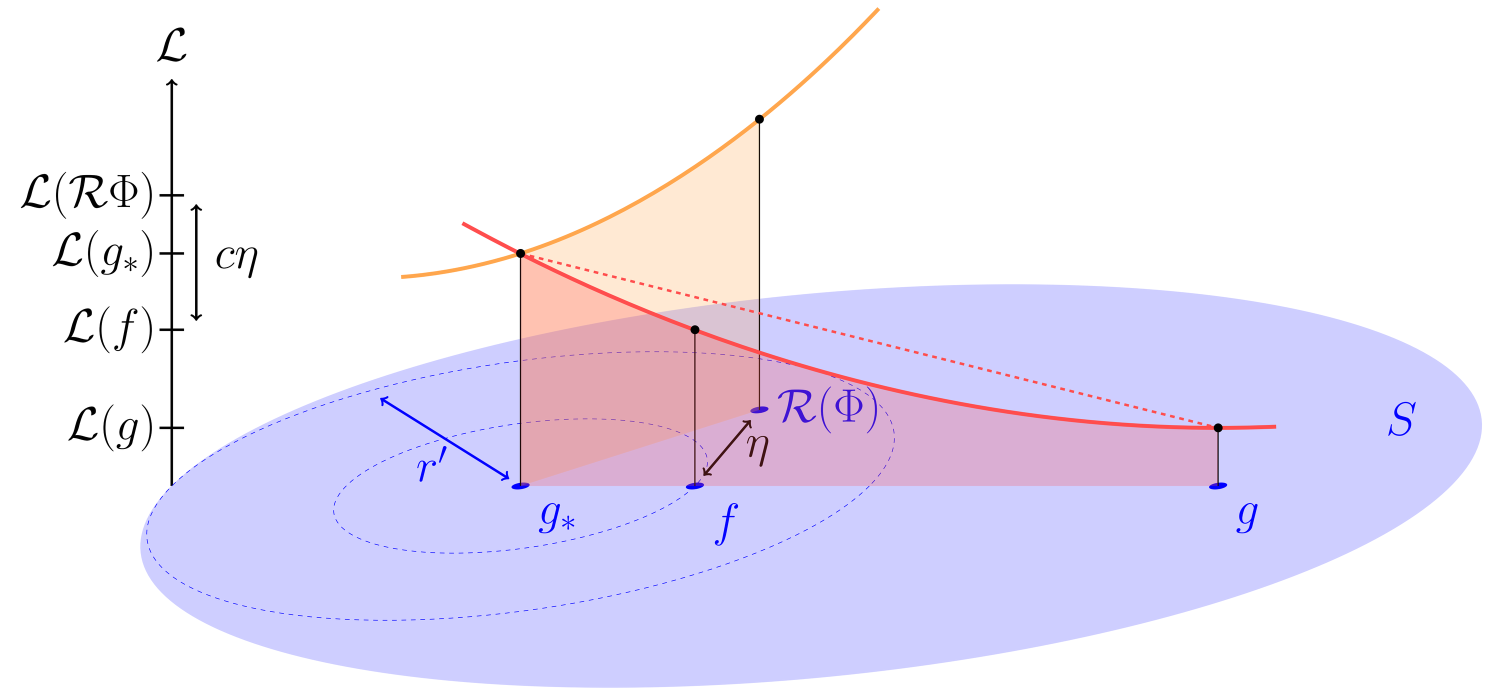
$$S := \{f \colon \Lambda(f) \le C\}$$

be compact in the $\|\cdot\|$-closure of $\bigcup_N \mathcal{R}(\mathcal{P}_N)$. We denote

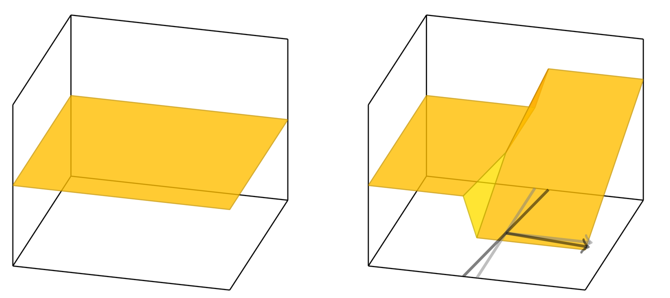
$$\Omega_N := \{\Phi \in \mathcal{P}_N \colon \Lambda(\mathcal{R}(\Phi)) \le C\}.$$

Then there is an architecture $N_{\varepsilon,r}$ such that for every local min. Γ_* of $\min_{\Gamma \in \Omega_{N_{\varepsilon,r}}} \mathcal{L}(\mathcal{R}(\Gamma))$ with radius at least r it holds that

$$\mathcal{L}(\mathcal{R}(\Gamma_*)) \le \min_{\Gamma \in \Omega_{N_{\varepsilon,r}}} \mathcal{L}(\mathcal{R}(\Gamma)) + \varepsilon.$$

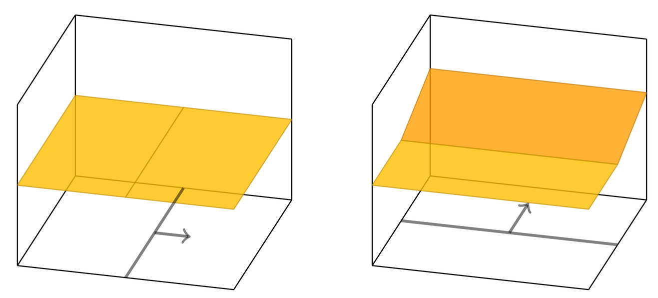


Failure - Exploding Gradient



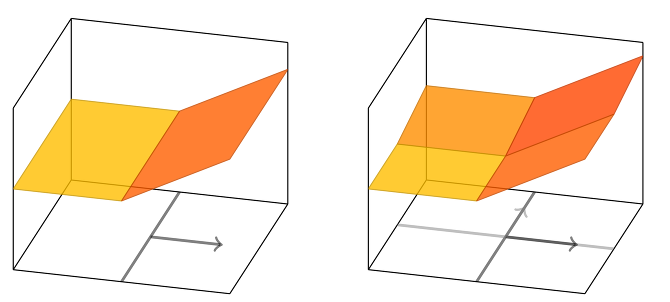
- $\Gamma := (0,0) \in \mathcal{N}_{(2,2,1)}$
- $g_k(x) := k\,\mathrm{ReLU}(\langle (k,0), x\rangle) - k\,\mathrm{ReLU}(\langle (k, -\frac{1}{k^2}), x\rangle)$

Failure - Unbalancedness



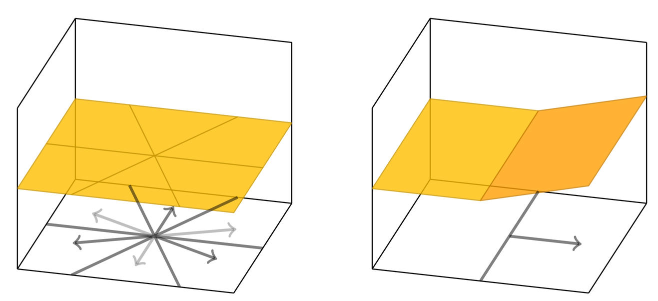
- $\Gamma := \big((r,0),0\big) \in \mathcal{N}_{(2,1,1)}$
- $g_k(x) := \frac{1}{k}\mathrm{ReLU}(\langle (0,1), x\rangle)$

Failure - Redundant Directions



- $\Gamma := \left(\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, (1,1)\right) \in \mathcal{N}_{(2,2,1)}$
- $g_k(x) := 2\,\mathrm{ReLU}(\langle (1,0), x\rangle) + \frac{1}{k}\mathrm{ReLU}(\langle (0,1), x\rangle)$

Failure - Opposite Weights



- $\Gamma := \big([A|-A]^T, (1,-1)\big) \in \mathcal{N}_{(d,2m,1)}$ with $\sum_{i=1}^m A_{:,i} = 0$
- $g_k(x) := \frac{1}{k}\mathrm{ReLU}(\langle v, x\rangle)$