Why neural networks find simple solutions:

the many regularizers of geometric complexity

Benoit Dherin
dherin@google.com

Mihaela Rosca
mihaelacr@deepmind.com

Michael Munn
munn@google.com

David Barrett
barrettdavid@deepmind.com

Google + DeepMind

BENOIT DHERIN
dherin@google.com

MICHAEL MUNN
munn@google.com

MIHAELA ROSCA
mihaelacr@deepmind.com

DAVID BARRETT
barrettdavid@deepmind.com

Why neural networks find simple solutions: the many regularizers of geometric complexity

Benoit Dherin*
Google
dherin@google.com

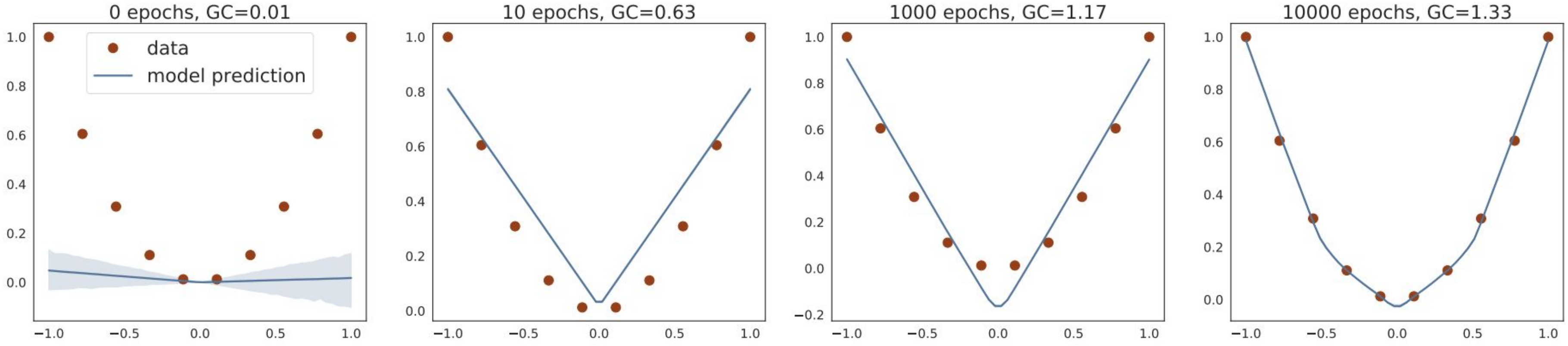
Michael Munn*
Google
munn@google.com

Mihaela C. Rosca
DeepMind, London
mihaelacr@deepmind.com

David G.T. Barrett
DeepMind, London
barrettdavid@deepmind.com

https://arxiv.org/abs/2209.13083

Motivating Example



Trained a ReLU MLP with 3 layers, 300 neurons each

The arc length of the learned function is minimized during training.

From arc length to Geometric Complexity

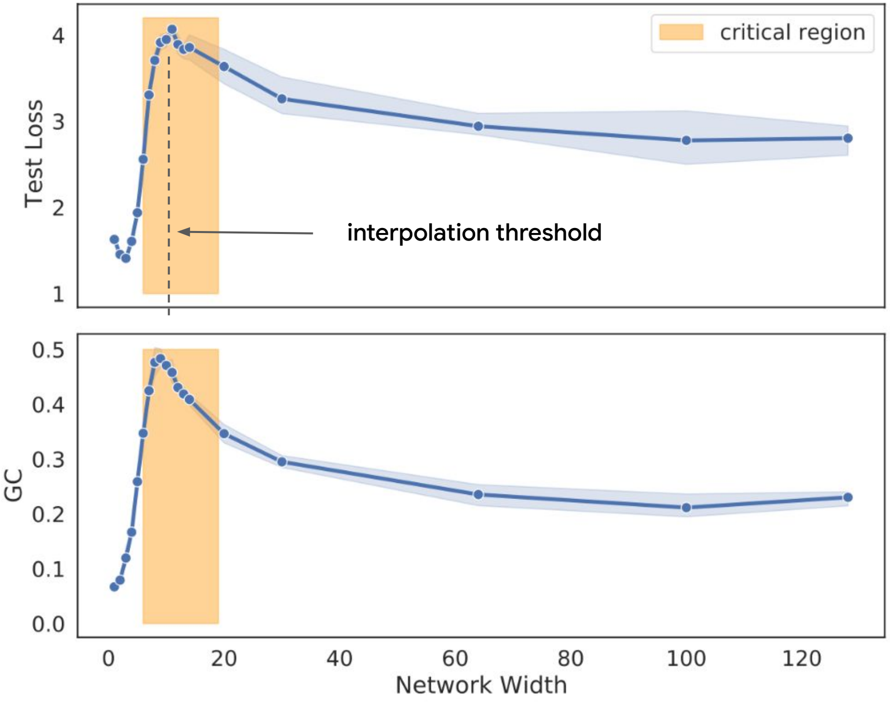
$$\text{arclength} = \int_{X_D} \sqrt{1 + \|\nabla_x f_\theta(x)\|_F^2}\, dx \simeq \int_{X_D} 1 + \frac{1}{2}\|\nabla_x f_\theta(x)\|_F^2\, dx = \text{Vol}(X_D) + \frac{1}{2}\int_{X_D} \|\nabla_x f_\theta(x)\|_F^2\, dx$$

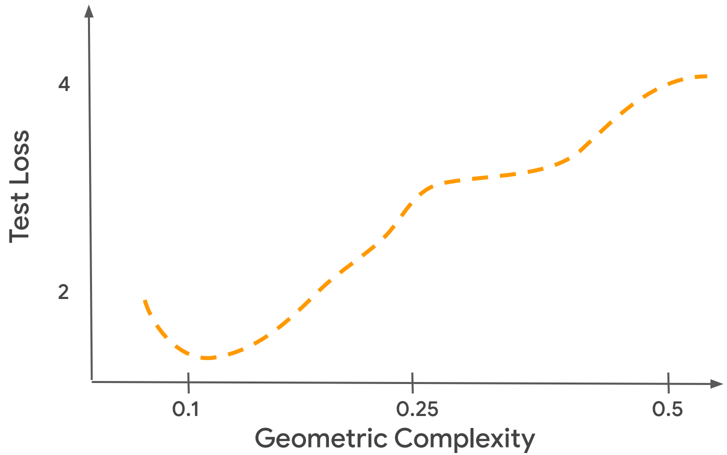
Geometric Complexity (GC)

$$\langle f_\theta, D\rangle_G = \frac{1}{|D|}\sum_{x\in D} \|\nabla_x f_\theta(x)\|_F^2$$

Discrete version of the Dirichlet Energy

GC captures the double-descent phenomenon

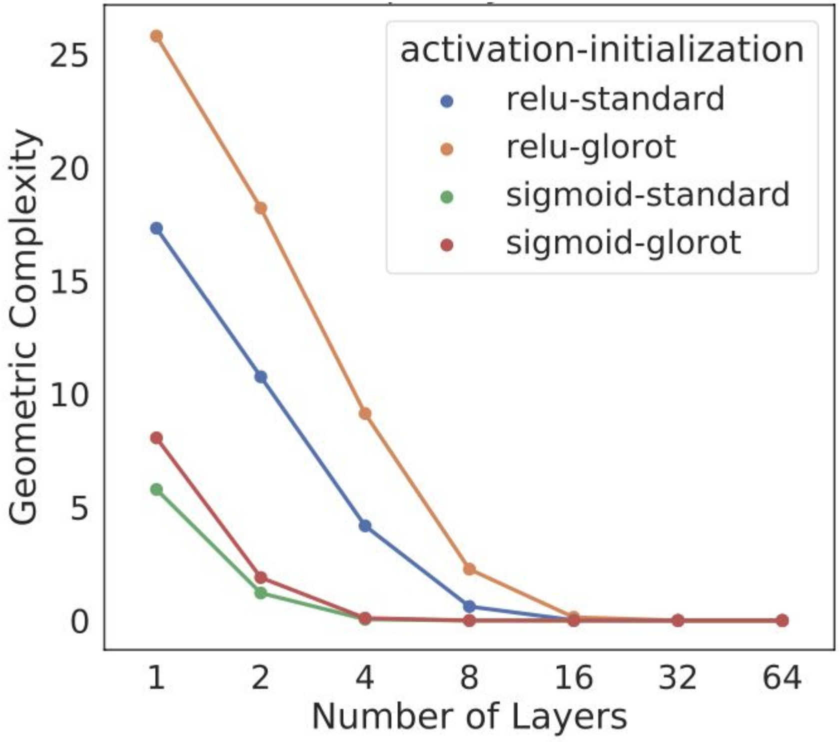




GC recovers the classical U-curve when used as the model complexity measure

Well-tuned neural networks find solutions with low geometric complexity

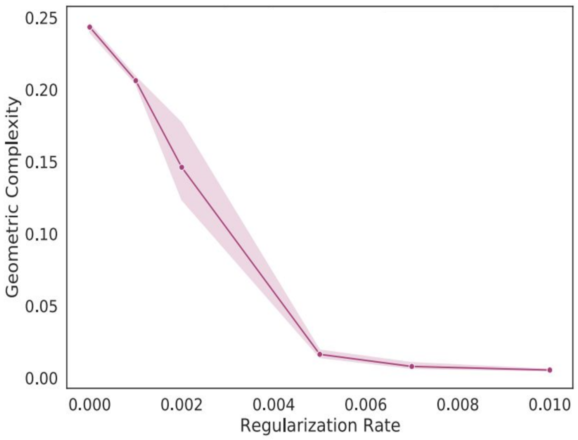
Deeper networks have lower GC at initialization



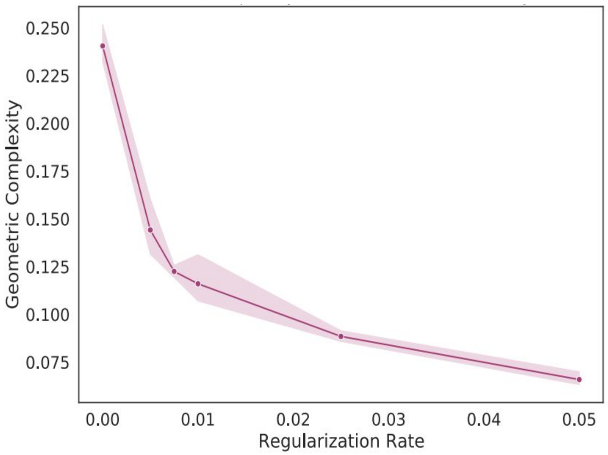
With deep enough neural networks, the model function is initialized to near the zero function and has lowest possible Geometric Complexity.

Common regularization schemes decrease GC

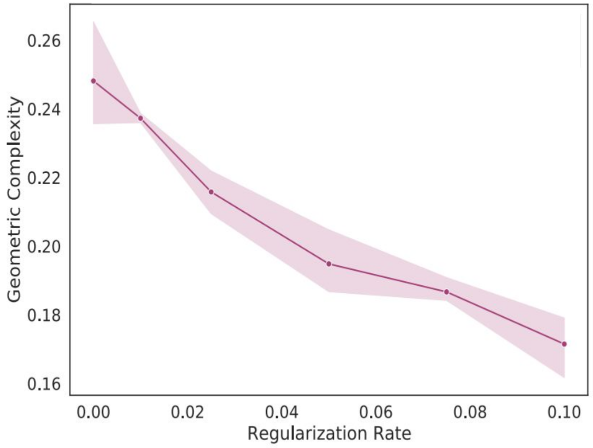
L2 Regularization



Flatness Regularization

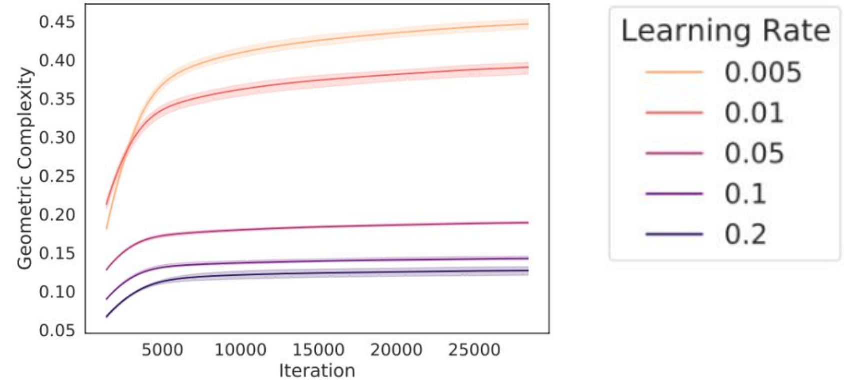


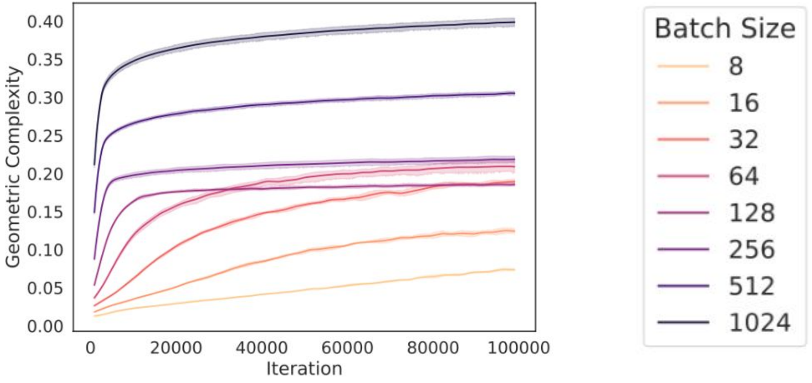
Spectral Regularization



Well regularized neural networks end up not only to be more performant but also simpler.

LR and batch size tuning decreases GC





IMPLICIT GRADIENT REGULARIZATION

David G.T. Barrett*
DeepMind
London
barrettdavid@google.com

Benoit Dherin*
Google
Dublin
dherin@google.com

The pressure of **implicit gradient regularization** transfers to a pressure on the geometric complexity.

For neural networks, SGD encourages simple solutions

Transfer Theorem. Consider a network $f_\theta : \mathbb{R}^d \to \mathbb{R}^k$ with ℓ layers parameterized by $\theta = (w_1, b_1, \ldots, w_l, b_l)$, then we have the following inequality

derivative w.r.t parameters

derivative w.r.t inputs

$$\|\nabla_x f_\theta(x)\|_F^2 \leq \frac{\|\nabla_\theta f_\theta(x)\|_F^2}{T_1^2(x,\theta) + \cdots + T_l^2(x,\theta)} \tag{43}$$

For well-tuned neural networks trained to minimal loss, the learned function is implicitly encouraged to find the most geometrically simple solution.

Google + DeepMind




BENOIT DHERIN
dherin@google.com

MICHAEL MUNN
munn@google.com

MIHAELA ROSCA
mihaelacr@deepmind.com

DAVID BARRETT
barrettdavid@deepmind.com

Why neural networks find simple solutions: the many regularizers of geometric complexity

Benoit Dherin[*]
Google
dherin@google.com

Michael Munn[*]
Google
munn@google.com

Mihaela C. Rosca
DeepMind, London
mihaelacr@deepmind.com

David G.T. Barrett
DeepMind, London
barrettdavid@deepmind.com

Thanks for listening

https://arxiv.org/abs/2209.13083