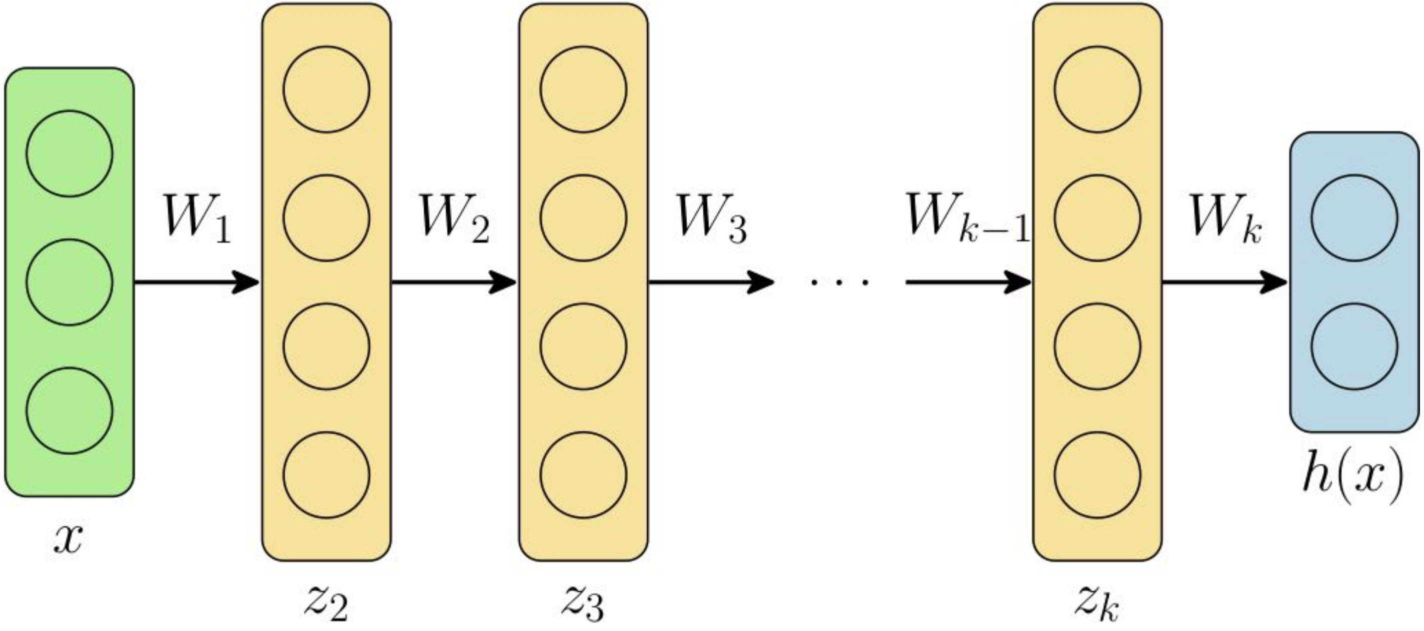
Deep equilibrium networks are sensitive to initialization statistics

Atish Agarwala and Sam Schoenholz

ICML 2022

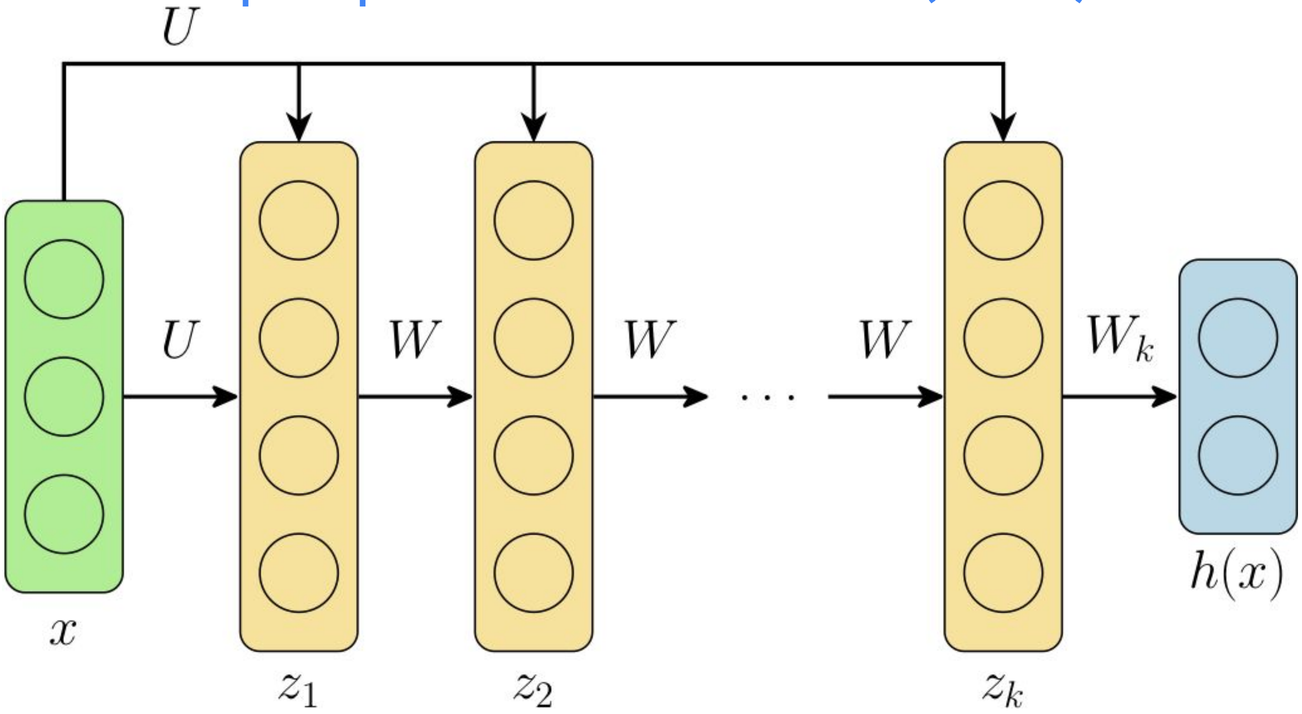
Google Research

What is a Deep Equilibrium Network (DEQ)?



Google Research

What is a Deep Equilibrium Network (DEQ)?



Google Research

How do DEQs differ from deep networks?

DEQs trade off *memory* (less weights) for *compute* (fixed point solving).

Are there other differences?

What are the *dynamical* effects of *reusing* parameters?

How should we think about *initializing* DEQs?

Paper outline

01 Theory of linear DEQs

02 Theory of non-linear DEQs

03 Initialization experiments

Linear DEQs

Linear DEQ:

$$\mathbf{z}^* = \mathbf{W}\mathbf{z}^* + \mathbf{x}$$

Explicit solution:

$$\mathbf{z}^* = (\mathbf{I} - \mathbf{W})^{-1}\mathbf{x}$$

Sensitive to **large eigenvalues** of **W**!

Expressivity requires ||**W**|| close to 1. If spectral norm ||**W**||>1, stability is lost.

What happens as ||W|| -> 1?

Google Research

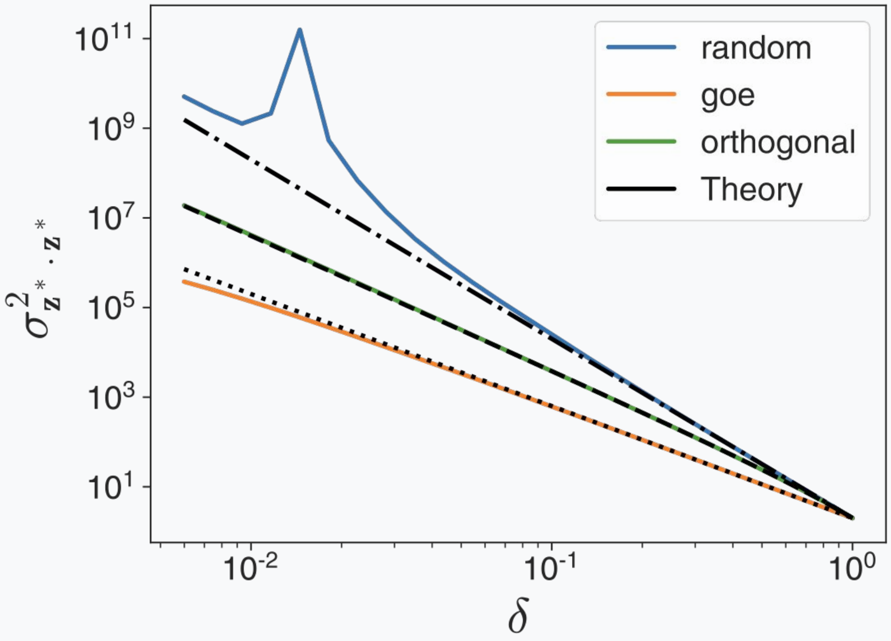
Random matrix families

DEQs are sensitive to higher order matrix statistics!

Families studied:

- **Random** - i.i.d. Gaussian entries.
- **Orthogonal** - random orthogonal matrices
- **GOE** - rotationally invariant family of symmetric matrices

Random family has more fluctuations than orthogonal or GOE!



Fluctuations in length of z* as ||W|| goes to 1.

Google Research

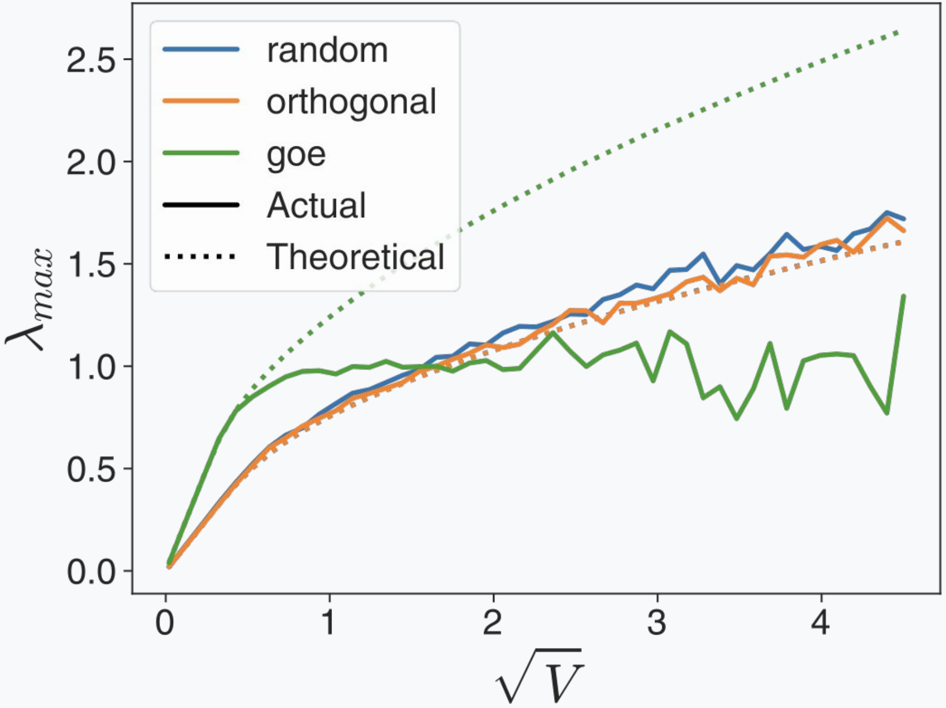
Non-linear DEQ

$$\mathbf{z}^* = \phi(\mathbf{W}\mathbf{z}^*) + \mathbf{x}$$

Analogous behavior to linear case!

Depends on Jacobian spectral norm ||**J**|| instead of ||**W**||.

Relationship between random, orthogonal, and GOE similar to linear case, see paper for details!



Experiments

Theory suggests that orthogonal and GOE initializations may provide stability!

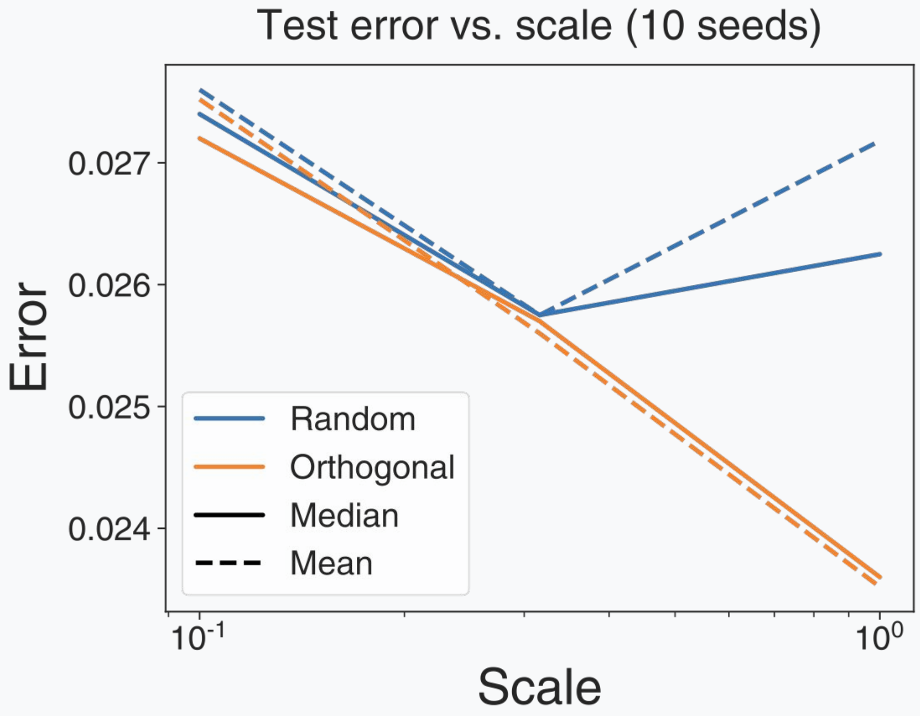
Studied fully connected DEQ on MNIST - theory covers this architecture.

Also conducted experiments on Wikitext-103 using DEQ-transformer.

DEQ FCN - MNIST

For a fully-connected DEQ on MNIST, orthogonal initializations outperform random initializations.

Reduced variance allows for larger initial weight matrices to be used, leading to better performance.



DEQ Transformer

Trained DEQ-Transformer architecture on wikitext-103.

Random has lowest best-case perplexity, but average-case performance plagued by training instability.

GOE ensures stability at some performance cost. Orthogonal interpolates between the two.

	GOE		Orthogonal		Random	
$\sqrt{V}$	Min	Ave	Min	Ave	Min	Ave
0.1	68.1	71.8	60.7	162.7	56.8	153.9
0.3	66.1	69.8	60.5	173.4	56.3	224.9
1.0	66.3	68.3	57.4	112.1	55.6	481.5

Conclusions

- Deep equilibrium networks are sensitive to higher-order statistics of weight matrices.
- Alternate initialization schemes (orthogonal, GOE) reduce variability as spectral norm goes to 1.
- Orthogonal initialization scheme can improve performance and trainability in practice.

Thanks for listening!