Fixed Points of Nonnegative Neural Networks

Tomasz Piotrowski

Talk based on:

TJ Piotrowski, RLG Cavalcante, M Gabor, Fixed points of nonnegative neural networks," Journal of Machine Learning Research, 25(139), 2024.

See also:

M Gabor, TJ Piotrowski, RLG Cavalcante, Positive concave deep equilibrium models, Proceedings of the 41st International Conference on Machine Learning, PMLR 235.

Motivation: nonnegativity

- Nonnegative inputs appear naturally in spectral analysis, image and text processing

- Equip neural networks with the ability of decomposing the input signal into additive sparse components, thus providing an understandable hierarchical representation of the input data

- Recent resurgence of analog neural networks, with weight parameters physically encoded in the conductance of programmable resistors

Motivation: fixed point analysis of neural networks

- For autoencoder networks, their fixed points are exactly the inputs that can be perfectly reconstructed

- A host of iterative methods can be cast as a fixed-point iteration, e.g., of the gradient descent-type:

$$x_{n+1} := (Id - \alpha \nabla f)(x_n), \quad n = 0, 1, 2, \ldots$$

- Application to deep equilibrium networks (DEQs): see our paper at ICML 2024

Standard neural network model

$$T \overset{\text{df.}}{=} T_n \circ \cdots \circ T_1$$

$$T_i(x_{i-1}) = \sigma_i(y_i) = \Big(\sigma_{i_1}(y_{i_1}), \sigma_{i_2}(y_{i_2}), \ldots, \sigma_{i_{k_i}}(y_{i_{k_i}})\Big)$$

$$y_i = W_i x_{i-1} + b_i$$

Standard neural network model

$$T \overset{\text{df.}}{=} T_n \circ \cdots \circ T_1$$

$$T_i(x_{i-1}) = \sigma_i(y_i) = \Big(\sigma_{i_1}(y_{i_1}), \sigma_{i_2}(y_{i_2}), \ldots, \sigma_{i_{k_i}}(y_{i_{k_i}})\Big)$$

$$y_i = W_i x_{i-1} + b_i$$

$T^F \overset{\text{df.}}{=} T_n \circ \cdots \circ T_1$ - EEG forward (compression) model

$T^I \overset{\text{df.}}{=} T_N \circ \cdots \circ T_{n+1}$ - EEG inverse (decompression) model

$T^{F\to I} \overset{\text{df.}}{=} T^I \circ T^F$ $\quad \mathrm{Fix}(T^{F\to I}) \overset{\text{df.}}{=} \{x \in \mathbb{R}^s : \ T^{F\to I}(x) = x\}$

Nonnegative networks with nonnegative biases are monotonic and (weakly) scalable

For nonnegative weights and biases and concave activation functions, the resulting neural network is monotonic and (weakly) scalable:

$$\forall x, \tilde{x} \in \mathbb{R}_+^k \quad x \leq \tilde{x} \implies f(x) \leq f(\tilde{x})$$

$$\forall x \in \mathbb{R}_+^k \ \forall \rho \geq 1 \quad f(\rho x) \leq \rho f(x)$$

$u \leq v \Leftrightarrow v - u \in \mathbb{R}_+^k$

Nonnegative networks with nonnegative biases are monotonic and (weakly) scalable

For nonnegative weights and biases and concave activation functions, the resulting neural network is monotonic and (weakly) scalable:

$$\forall x, \tilde{x} \in \mathbb{R}^k_+ \quad x \leq \tilde{x} \implies f(x) \leq f(\tilde{x})$$

$$\forall x \in \mathbb{R}^k_+ \ \forall \rho \geq 1 \quad f(\rho x) \leq \rho f(x)$$

under mild assumptions (e.g., sigmoid used or positive weights used at a single layer):

$$\forall x \in \mathbb{R}^k_+ \ \forall \rho > 1 \quad f(\rho x) \ll \rho f(x)$$

$u \ll v \Leftrightarrow v - u \in \mathrm{int}(\mathbb{R}^k_+)$

Spectral radius of a monotonic and (weakly) scalable network

Asymptotic mapping:

$$T_\infty : \mathbb{R}_+^k \to \mathbb{R}_+^k : x \mapsto \lim_{p\to\infty} \frac{1}{p} T(px)$$

Spectral radius:

$$\rho(T_\infty) = \max\{\lambda \in \mathbb{R}_+ : \ \exists\, x \in \mathbb{R}_+^k \backslash \{0\} \text{ s.t. } T_\infty(x) = \lambda x\} \in \mathbb{R}_+$$

Spectral radius of a monotonic and (weakly) scalable network

Asymptotic mapping:

$$T_\infty : \mathbb{R}_+^k \to \mathbb{R}_+^k : x \mapsto \lim_{p\to\infty} \frac{1}{p} T(px)$$

Spectral radius:

$$\rho(T_\infty) = \max\{\lambda \in \mathbb{R}_+ : \ \exists\, x \in \mathbb{R}_+^k \backslash \{0\} \text{ s.t. } T_\infty(x) = \lambda x\} \in \mathbb{R}_+$$

We have the following result:

$$T_\infty(x) = W_n \ldots W_2 W_1(x) \text{ or } T_\infty(x) = 0$$

Fixed points and spectral radius

If a neural network $T\colon \mathbb{R}_+^k \to \mathbb{R}_+^k$ is monotonic and weakly scalable and grows fast enough* and $\rho(T_\infty) < 1$, then $\mathrm{Fix}(T)$ is an interval.

*Is upper- and lower-primitive at its fixed points.

Fixed points and spectral radius

[1] If a neural network $T\colon \mathbb{R}_+^k \to \mathbb{R}_+^k$
is monotonic and scalable,
then T has a fixed point
if and only if $\rho(T_\infty) < 1$.
This fixed point is unique and positive.

[1] R. L. G. Cavalcante, Q. Liao, and S. Stańczak, “Connections between spectral properties of asymptotic mappings and solutions to wireless network problems,” IEEE Transactions on Signal Processing, vol. 67, no. 10, pp. 2747–2760, 2019.

Nonnegative monotonic networks

Consider a nonnegative and monotonic neural network, e.g., such that:

- weight operators are nonnegative,

- if a layer admits negative biases, then its activation function should be globally nonnegative and monotonic (e.g., composed of ReLU or sigmoid),

- if a layer admits an activation function which is only monotonic on $\mathbb{R}_+^k$ (e.g., composed of Swish, Mish, GELU), then its biases should be nonnegative.

Nonnegative monotonic networks

Consider a nonnegative and monotonic neural network, e.g., such that:

- weight operators are nonnegative,

- if a layer admits negative biases, then its activation function should be globally nonnegative and monotonic (e.g., composed of ReLU or sigmoid),

- if a layer admits an activation function which is only monotonic on $\mathbb{R}_+^k$ (e.g., composed of Swish, Mish, GELU), then its biases should be nonnegative.

For such a network, the fixed point iteration starting at $x_0 = 0$ converges to its least fixed point if it exists.

Performance results

Spectral radiuses, products of weight matrices spectral norms, spectral norms of products of weight matrices, and test loss on the MNIST dataset for different configurations of autoencoders.

Configuration	Test loss	$\rho(T_\infty)$	$\|W_1\| \cdot \|W_2\|$	$\|W_2 \cdot W_1\|$
Sigmoid NN	0.2308	0.00	7.271	6.052
Tanh NN	0.0065	0.00	463.9	187.3
Tanh PN	0.0255	0.00	40.63	35.85
ReLU spectral NN	0.0063	0.98	1.072	0.999
ReLU spectral PN	0.0059	0.99	1.0006	1.0001
Tanh + Swish NN	0.0041		15.221	8.8120
ReLU + Sigmoid NR	0.0025		2845.3	215.13
ReLU + Sigmoid RR	0.0052		3196.9	332.97

Performance results

Spectral radiuses, products of weight matrices spectral norms, spectral norms of products of weight matrices, and test loss on the MNIST dataset for different configurations of autoencoders.

Configuration	Test loss	$\rho(T_\infty)$	$\|W_1\| \cdot \|W_2\|$	$\|W_2 \cdot W_1\|$
Sigmoid NN	0.2308	0.00	7.271	6.052
Tanh NN	0.0065	0.00	463.9	187.3
Tanh PN	0.0255	0.00	40.63	35.85
ReLU spectral NN	0.0063	0.98	1.072	0.999
ReLU spectral PN	0.0059	0.99	1.0006	1.0001
Tanh + Swish NN	0.0041		15.221	8.8120
ReLU + Sigmoid NR	0.0025		2845.3	215.13
ReLU + Sigmoid RR	0.0052		3196.9	332.97

Open questions & Next steps

1. How robust is the reconstruction of points which are *approximate* fixed points?
2. How can nonnegative networks be trained efficiently?
3. What is the shape of the fixed point set of generic nonnegative monotonic neural networks?

4. How to expand expressibility of nonnegative neural networks while maintaining their benefits?

Thank you for your attention