Growth and harvest induce essential dynamics in neural networks

Ilona Kulikovskikh
Samara University
Samara, Russia
kulikovskikh.im@ssau.ru

Tarzan Legović
Institute of Applied Ecology, Oikon Ltd.
Libertas International University
Ruđer Bošković Institute
Zagreb, Croatia
tlegovic@oikon.hr

ABSTRACT

Training neural networks with faster gradient methods brings them to the edge of stability, proximity to which improves their generalization capability. However, it is not clear how to stably approach the edge. We propose a new activation function to model inner processes inside neurons with single-species population dynamics. The function induces essential dynamics in neural networks with a growth and harvest rate to improve their generalization capability.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning*; *Bio-inspired approaches*;

KEYWORDS

neural networks, generalization, stability, population dynamics

ACM Reference Format:
Ilona Kulikovskikh and Tarzan Legović. 2021. Growth and harvest induce essential dynamics in neural networks. In *2021 Genetic and Evolutionary Computation Conference Companion (GECCO '21 Companion), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3449726.3459421

1 INTRODUCTION

When trained with gradient-based methods, a neural network is expected to meet two criteria: converge faster and generalize better [1, 5]. While accelerating convergence involves reducing the model errors on a training dataset, good generalization capability requires minimizing the difference between the network errors on a training and a testing dataset, which plays a crucial role.

Addressing the generalization issue, Trask et al.[11] pointed to the importance of a network ability to extrapolate to the range of values unseen during training. Wang et al. [13] showed that deep sequence learning models fail to generalize on testing data due to distribution shifts from dynamic system chaotic behavior. Looking at the problem from the perspective of adaptive optimization, recent studies [3, 4, 9] revealed that a step size, maximizing the test accuracy, is usually larger than a step size minimizing the training loss. The chaotic nature of faster gradient methods demystifies this phenomenon [12]. A larger step size brings the optimizer to the edge of stability [3, 4], training sufficiently close to which, according to the concept of chaos [10], increases generalization capability. But, how to choose a step size to reach this edge remains unclear.

We refer to single-species population dynamics [7, 8] to regulate this process explicitly and to induce essential dynamics in neural networks. A new non-monotonic activation function is built on an original s-shaped monotonic function (sigmoid) but exhibits more complex behavior. It is equiped with the growth and harvesting rates to self-stabilize the model dynamics close to the edge of stability and, thus, to increase its generalization capability.

GECCO '21 Companion, July 10–14, 2021, Lille, France

ACM ISBN 978-1-4503-8351-6/21/07.
https://doi.org/10.1145/3449726.3459421

2 NETWORK GROWTH AND HARVEST

We introduce a new activation function - LIGHT (**L**og**I**stic **G**rowth with **H**arves**T**ing) - to model inner processes inside neurons. It enriches the state-of-the-art perspective of how neurons receive electrical impulses from other cells, accumulate them, and generate an action potential spike if a threshold value is exceeded. The population of impulses starts growing with the rate r by the logistic law. After time T, it is harvested with the rate E. The sizes of population at $t = 0$ and $t = T$ are specified.

Definition. *For any time $t \in \mathrm{R}$, harvest time instant $T > 0$,* per capita *growth rate $r > 0$ and* per capita *harvesting rate $E \geq 0$, a population of impulses inside a neuron $\ell^{r,E}(t)$, such that $\lim_{t\to-\infty} \ell^{r,E}(t) = 0$, $\lim_{t\to\infty} \ell^{r,E}(t) = \varepsilon$, where ε is the extent to which r is impacted by E, develops according to:*

$$\ell^{r,E}(t) = \varepsilon \mathrm{e}^{\left(\ln_{\mathrm{q}} N_T + \mathbf{1}_T(t)\frac{E}{r}\right)\mathrm{e}^{-r(t-T)}} \tag{1}$$

where a population size at T is $N_T = \ell^{r,E}(T)$ and $\ln_{\mathrm{q}}(x)$ is the q*-logarithm, where* q *is the rate with which population grows when smaller.*

The parameter q generalizes the Verhulst ($q = 1$) and Gompertz ($q \to 0$) laws of population dynamics. If $q = 1$, $T = 0$, $N_0 = 0.5$, $r = 1$, and $E = 0$, the function reduces to the sigmoid.

We adopt this function to minimize an empirical loss function for any dataset $\{\mathrm{x}_i, y_i\}_{i=1}^m$ with $\mathrm{x}_i \in \mathrm{R}^n$, $y_i \in \{-1, 1\}$, for each mini-batch subset $B(t) \subseteq \{1, \ldots, m\}$ with a weight vector $\Theta \in \mathrm{R}^n$:

$$\mathcal{L}^{r,E}(\Theta) = \sum\nolimits_{i\in B(t)} \ell^{r,E}\left(y_i \langle \Pi(\Theta), \mathrm{x}_i\rangle\right), \tag{2}$$

where $\ell^{r,E}$ measures the discrepancy between the output y and the model prediction, $\Pi(\Theta) = \Theta_1 \times \Theta_2 \times \cdots \times \Theta_L$, $\Theta = \{\Theta_l \in \mathrm{R}^{d_{l-1}\times d_l} : l = 1, 2, \ldots, L\}$, L is the number of layers, d_l is the number of nodes in the layer l. While solving (2) with a non-adaptive

SGD optimizer, the function (1) stably increases a fixed step size η with regard to a growth rate r and a harvesting rate E.

3 RESULTS

To examine the impact of inducing growth and harvest, we implemented a neural network with a hidden layer $L = 1$ of the ReLU neurons $d_l = 5$, where the proposed function is applied only to the output. We generated a set of synthetic datasets ($m = 1000$, $n = 2$), which were randomly split into training (80%) and testing (20%) subsets. The samples were equally divided between the classes. We compared non-adaptive methods - SGD with the sigmoid (**s-sgd**) and SGD with the proposed function (**light-sgd** - to two popular adaptive methods with the sigmoid - Adam (**s-adam**) and AdaGrad (**s-adagrad**). For all the optimizers, we used the default parameters and batch size $|B(t)| = 75$. The proposed function was implemented as a custom output activation layer with Keras `class LIGHT(Layer)`.

Table 1 shows that the optimizer with induced growth and harvesting rates significantly outperforms the other methods. The best values of the test accuracy (mean± std) averaged over $n_{\text{run}} = 10$ and $n_{\text{epoch}} = 1500$ are highlighted in bold. The best balance between train and test accuracy over epochs can be seen in Figure 1, upper triangle of the plots).

Table 1: Test accuracy on synthetic datasets (mean± std, %)

dataset	s-adam	s-adagrad	s-sgd	light-sgd
Fig. 1 (a)	99.64±0.36	99.54±0.5	99.68±0.36	**99.98±0.021**
Fig. 1 (b)	96.63±0.52	96.44±0.94	96.67±0.52	**96.62±0.5**
Fig. 1 (c)	98.21±1.16	99.24±0.76	95.38±9.43	**99.63±0.34**
Fig. 1 (d)	88.14±6.6	87.2±8.47	89.98±1.62	**91.7±1.16**
Fig. 1 (e)	97.12±1.8	91.2±16.76	95.93±5.54	**99.86±0.2**
Fig. 1 (f)	84.98±2.47	85.28±2.6	83.45±6.49	**86.74±1.28**
Fig. 1 (g)	90.61±2.62	94.28±4.42	89.37±2.22	**98.48±2.9**
Fig. 1 (h)	87.2±2.02	89.92±3.42	86.25±1.62	**93.96±0.76**

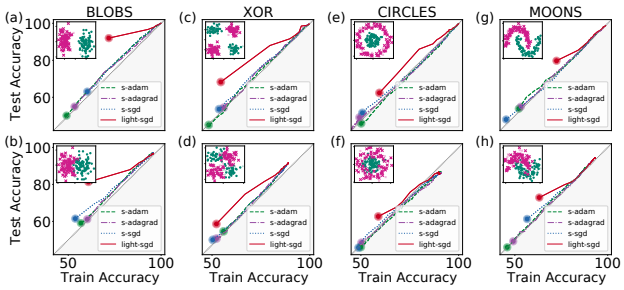

Figure 1: Test/train accuracy curves over epochs

The success of **light-sgd** can be attributed to the self-stabilization of the model dynamics: while an increase in r pushes the system towards not clearly defined edge of stability, a simultaneous increase in E fixes this edge resulting in higher generalization capability. As we can see, this effect becomes less apparent with iterations and, thus, requires building the communities of the LIGHT neurons to enhance the revealed dynamics.

We applied the Wilcoxon signed-rank test to the mean and std of accuracy curves, averaged over $n_{\text{run}} = 10$, to evaluate the differences in performance between the methods. Since p-values turned out to be less than the 0.01 significance level, we rejected the null hypothesis and concluded that the presented results are statistically significant. The proposed function was also validated on MNIST, Fashion MNIST, and CIFAR10 datasets (see Table 2). The labels of the image classification datasets were binarized with the target class {5}. The samples were randomly extracted ($m = 1000$) from each of them and split into training (80%) and testing (20%) subsets. To classify the images, we used the pre-defined growth and harvesting rate: $r = 4.08$, $E = 6.4$.

Table 2: Test accuracy on image datasets (mean± std, %)

dataset	s-adam	s-adagrad	s-sgd	light-sgd
MNIST	96.43±1.06	96.71±0.84	96.72±0.84	**96.83±0.54**
F. MNIST	96.99±1.29	97.14±0.89	97.15±1.02	**97.24±0.75**
CIFAR10	87.25±2.27	84.08±2.28	87.44±2.51	**89.6±1.5**

4 CONCLUSION AND DISCUSSION

We proposed a new activation function to model inner processes inside neurons with single-species population dynamics and demostrated that growth and harvesting rates induce essential dynamics in neural networks and, thus, improve their generalization capability. The study supports the idea of De Felice et al. [6] that the biological activation function has a more complicated behavior which reduces to the step or sigmoid function for some hyperparameters describing its shape. Also, shapes of the proposed function share similarities with evolving activations discussed in [2].

ACKNOWLEDGMENTS

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions.

REFERENCES

[1] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. 2019. A convergence analysis of gradient descent for deep linear neural networks. In *ICLR*.
[2] Garrett Bingham, William Macke, and Risto Miikkulainen. 2020. Evolutionary optimization of deep learning activation functions. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference (GECCO '20)*. 289–296.
[3] Valentin De Bortoli, Alain Durmus, Xavier Fontaine, and Umut Simsekli. 2020. Quantitative Propagation of Chaos for SGD in Wide Neural Networks. In *NeurIPS*.
[4] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. 2021. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. In *ICLR*.
[5] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. 2017. Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression. *Journal of Machine Learning Research* 18 (2017), 1–51.
[6] P. De Felice, C. Marangi, G. Nardulli, G. Pasquariello, and L. Tedesco. 1993. Dynamics of neural networks with non-monotone activation function. *Network: Computation in Neural Systems* 4, 1 (1993), 1–9.
[7] William G. Gray and Genetha A. Gray (Eds.). 2017. *Introduction to Environmental Modeling*. Cambridge University Press, Cambridge, UK.
[8] T. Legović. 2016. Dynamic population models. In *Ecological model types*, S.E. Jorgensen (Ed.). Elsevier, 39–63.
[9] Zhiyuan Li and Sanjeev Arora. 2020. An Exponential Learning Rate Schedule for Deep Learning. In *ICLR*.
[10] Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. 2017. Deep information propagation. In *ICLR*.
[11] Andrew Trask, Felix Hill, Scott Reed, Jack Rae, Chris Dyer, and Phil Blunsom. 2018. Neural Arithmetic Logic Units. In *NeurIPS*.
[12] Kees van den Doel and Uri Ascher. 2012. The Chaotic Nature of Faster Gradient Descent Methods. *Journal of Scientific Computing* 51 (2012), 560–581.
[13] Rui Wang, Danielle Robinson, Christos Faloutsos, Yuyang Wang, and Rose Yu. 2020. Learning dynamical systems requires rethinking generalization. In *1st NeurIPS workshop on Interpretable Inductive Biases and Physically Structured Learning*.