Incorrect Application of Yilmaz–Poli (2022) Initialisation Method in dePater–Mitici 2023 paper entitled "A mathematical framework for improved weight initialization of neural networks using Lagrange multipliers"

Riccardo Poli[a], Ahmet Yilmaz[b,*]

[a]*School of Computer Science and Electronic Engineering, University of Essex, Colchester, CO4 3SQ, UK*
[b]*Department of Computer Engineering, Karamanoglu Mehmetbey University, Karaman, Turkey*

Abstract

In this letter to the editor we report on a methodological error made in the article entitled "A mathematical framework for improved weight initialization of neural networks using Lagrange multipliers" by dePater and Mitici recently appeared in this journal. The error relates to the incorrect application of a weight initialisation method we derived, published last year in this same journal.

1. Background

In [1], de Pater and Mitici presented an effective methodology ("Lagrangian LR") to initialise the last layer of a neural network to improve performance. To prove the effectiveness of their method, they tested it on one regression problem and one classification problem (CIFAR-100), using two different network architectures (ResNet-18 and ResNet-34) for the latter. They compared the performance obtained when all layers of a network were initialised with a variety of previously-published initialisation methods *vs* when the last layer was, instead, initialised with *Lagrangian LR*.

One of the previously-published initialisation methods used for this purpose was the one we proposed in [2] as a potential solution for the vanishing gradient problem in deep neural networks with logistic activation function.

While we are very grateful that our method was considered by de Pater and Mitici for the evaluation of their excellent technique, we believe the method was used in an erroneous manner (more in the next section), which resulted in some of the results reported in their paper providing an unfair representation of the quality of our initialisation technique.

2. Methodological Error

Key features of our method [2] is that it was *theoretically derived for the logistic activation function* and it prescribes that in networks using such a function, *the weights should be initialised with a specific negative mean* (with the usual standard deviation).

Unfortunately, the networks utilised in [1] *did not use a logistic activation function.* For this reason, the method was not applicable and the results reported in that article cannot be considered a representation of the method's quality or performance. More details are provided in the next two sub-sections.

*Corresponding author
Email addresses: rpoli@essex.ac.uk (Riccardo Poli), yilmazahmet@kmu.edu.tr (Ahmet Yilmaz)

2.1. Results for [1]'s Regression Problem

In [1]'s regression problem a network with a *hyperbolic-tangent activation functions* was used. In this case, the use of a negative mean for the initial weights did not provide any speed up in the training, but it also did not cause any major issues.

Indeed, in [1, Table 1], networks initialised with our method ended up having the same test error as those initialised with the Lagrangian LR method proposed by de Pater and Mitici.

2.2. Results for [1]'s Classification Problem

Much more severe were the consequences of the incorrect application of our initialisation method on the classification results. As mentioned above, for the classification problem (CIFAR), de Pater and Mitici adopted ResNets which use the *RELU activation function*.

Generally RELU neurons suffer from the *dying neurons problem*: if there is a large gradient for a connection weight, when the weight is updated, the corresponding neuron may become sufficiently inhibited to output zero thereafter (or die). If a neuron dies, it can never be active again. As shown in recent theoretical work [3], with traditional, zero-mean, weight initialisation, a significant proportion of the neurons of a RELU network are "born dead" (i.e., they already are in a permanent off state), the solution being proposed being to use an asymmetric, *positive-mean*, distribution instead.

It stands to reason that the dead/dying neurons problem is exacerbated if one initialises the weights with a negative mean as was done by applying our method to initialise the ResNets. One does not need to do complex theory to see this. With standard input scaling, the input patterns in the CIFAR problem will have positive values. When such positive inputs are fed into the first layer of RELU neurons and the connection weights are drawn from a distribution with negative mean, most likely the corresponding net inputs will be negative. So, one should expect most such neurons to be dead or near dead (i.e., active only on a small subset of input patterns). Of course, this, on its own, would severely limit what the following layers could achieve. However, such layers, too, are hampered by a worsened dead neurons problem, as their inputs are non-negative (being the outputs of RELU functions) but the weights have negative mean.

We strongly believe that the described *exacerbation of the dying/dead neurons problem* caused by the improper application of our initialisation method is the reason why in the classification problem [1, Tables 3 and 4] the method scored poorly in comparison with other methods.

3. Conclusion

While we have no doubts on the effectiveness of the method report in [1], the application of our initialisation method [2] in that article was incorrect. Our method should not have been used to initialise networks using RELU nodes, and in principle, also with hyperbolic tangent nodes.

As a result of the incorrect application of our method, [1] gives the impression that our technique (and its complex theoretical derivation) is not effective.

For this reason, we respectfully request, that a correction of [1] is published which either makes no reference to our method or it uses in networks with logistic activation function.

References

[1] I. de Pater, M. Mitici, A mathematical framework for improved weight initialization of neural networks using lagrange multipliers, Neural Networks 166 (2023) 579–594.

[2] A. Yilmaz, R. Poli, Successfully and efficiently training deep multi-layer perceptrons with logistic activation function simply requires initializing the weights with an appropriate negative mean, Neural Networks 153 (2022) 87–103.

[3] L. Lu, Y. Shin, Y. Su, G. E. Karniadakis, Dying ReLU and initialization: Theory and numerical examples, arXiv preprint arXiv:1903.06733 (2019).