Notes on the paper “Flat Channels to Infinity in Neural Loss Landscapes” by Flavio Martinelli

Compiled by D. Gueorguiev, 9/1/25

# Introductory notes

In the evolving field of deep learning, researchers continue to uncover hidden dynamics in the training of neural networks that challenge traditional assumptions; the study shows that seemingly flat valleys in neural loss landscapes may hide “wormhole-like channels” where neurons’ weights diverge toward infinity, leading them to fuse into gated linear units (GLUs) and demonstrating that even simple fully connected layers can spontaneously develop gating mechanisms; this suggests that optimization is not just about finding minima but can induce emergent functional shifts, with potential business impacts ranging from unexpected model capabilities to new cost-performance tradeoffs in AI-driven products, though the effect remains mostly theoretical and observed in limited regression settings, meaning enterprises should view it as an intriguing signal of emergent design rather than a replacement for deliberate architectural innovation.

The loss landscapes of neural networks contain minima and saddle points that may be connected in flat regions or appear in isolation. The authors of this paper identify and characterize a special structure in the loss landscape: channels along which the loss decreases extremely slowly, while the output weights of at least two neurons,  and , diverge to , and their input weight vectors,  and , become equal to each other. At convergence, the two neurons implement a gated linear unit:

Geometrically, these channels to infinity are asymptotically parallel to symmetry-induced lines of critical points. Gradient flow solvers, and related optimization methods like SGD or ADAM, reach the channels with high probability in diverse regression settings, but without careful inspection they look like flat local minima with finite parameter values. The paper characterizes these quasi-flat regions in terms of gradient dynamics, geometry, and gives functional interpretation. The emergence of gated linear units at the end of the channels highlights a surprising aspect of the computational capabilities of fully connected layers.

A diagram of a graph

AI-generated content may be incorrect.

# References

[1] [Flat Channels to Infinity in Neural Loss Landscapes, Flavio Martinelli et al, EPFL, 2025](https://github.com/dimitarpg13/deep_learning_and_neural_networks/blob/main/literature/articles/Flat_Channels_to_Infinity_in_Neural_Loss_Landscapes_Martinelli_2025.pdf)

[2] [Convex Optimization, Steven Boyd, Lieven Vandenberghe, 2009](https://github.com/dimitarpg13/optimization_classification_regression/blob/main/literature/books/ConvexOptimization_Boyd_2004.pdf)

[3] [Stochastic Gradient Descent, Convex Optimization 10-725, CMU, slides, Ryan Tibshirani](https://github.com/dimitarpg13/statistical_learning_and_kernel_methods/blob/main/literature/articles/gradient_descent/stochastic-gradient-descent_Ryan_Tibshirani_slides.pdf)

[4] [Gating Mechanism, Wikipedia](https://en.wikipedia.org/wiki/Gating_mechanism)

[5] [Gated Recurrent Unit, Wikipedia](https://en.wikipedia.org/wiki/Gated_recurrent_unit)

# Appendix

## Gating mechanism in Long Short-Term Memory

The primary reason for the gating mechanism is to mitigate the vanishing gradient problem which is observed with the vanilla RNNs.