Piecewise linear activations substantially shape the loss surfaces of neural networks

Fengxiang He*, Bohan Wang*, Dacheng Tao

Problem:

- What does the loss surface of neural network look like?
- Does the loss surface of a nonlinear neural network different from linear ones?
- How would piecewise linear activations shape the loss surface?

Why they are Important?

- Optimization. The geometrical structure of the loss surface helps design innovative optimization methods especially when the landscape is highly non-convex and non-smooth.
- Generalization. Many works have established the relationship between the generalization ability and the sharpness/flatness of the local minima. The understanding of the landscape helps analyze the sharpness/flatness of the minima of neural network, and further study the generalization of deep learning.
- Estimation. Understanding the landscape of the loss surface can also help understand the estimation ability of deep learning.

Part 1:

- Neural networks with arbitrary depth and arbitrary piecewise linear activations (excluding linear functions) have infinitely many spurious local minima under arbitrary continuously differentiable loss functions.
- This result only relies on three mild assumptions that cover most practical circumstances: (1) the training sample set is linearly inseparable; (2) all training sample points are distinct; and (3) the output layer is narrower than the other hidden layers.

Proof idea:

1. Yun et al. (2019) prove that neural networks with one hidden layer and two-piece linear activations have spurious local minima.
2. We extend the conditions to neural networks with arbitrary hidden layers and two-piece linear activations.
3. We further extend the conditions to neural networks with arbitrary depth and arbitrary piecewise linear activations. Since some parameters of the constructed spurious local minima are from continuous intervals, we have obtained infinitely many spurious local minima.

Part 2:

- The loss surface of any nonlinear neural network is partitioned into smooth and multilinear open cells.
- Every local minimum is globally minimal within a cell.
- Local minima (probably including global minima) are connected within a cell by a continuous path, on which all points have the same empirical risk.
- The points on a local minimum valley are in an equivalent class. All the local minimum valleys are in a quotient space.
- Linear neural networks are covered by our theories as a simplified case.

Proof idea:

1. We prove that within every cell, the empirical risk $\widehat{\boldsymbol{R}}$ is convex with respect to a variable $\widehat{W}$ mapped from the weights W. So, all the local minima w.r.t. $\widehat{W}$ are also the global minima in the cell.
2. We prove that the local optimality is maintained under the constructed mapping. Specifically, the local minima of the empirical risk $\widehat{\boldsymbol{R}}$ w.r.t. W are also the local minima w.r.t $\widehat{W}$.

Fengxiang He, Bohan Wang, and Dacheng Tao, “Piecewise linear activations substantially shape the loss surfaces of neural networks”, ICLR, 2020.