Evaluating Loss Landscapes from a Topology Perspective

Tiankai Xie[*]
Arizona State University
txie21@asu.edu

Caleb Geniesse[*]
Lawrence Berkeley National Lab
cgeniesse@lbl.gov

Jiaqing Chen[*]
Arizona State University
jchen501@asu.edu

Yaoqing Yang
Dartmouth College
yaoqing.yang@dartmouth.edu

Dmitriy Morozov
Lawrence Berkeley National Lab
dmorozov@lbl.gov

Michael W. Mahoney
ICSI, LBNL, and UC Berkeley
mmahoney@stat.berkeley.edu

Ross Maciejewski
Arizona State University
rmacieje@asu.edu

Gunther H. Weber
Lawrence Berkeley National Lab
ghweber@lbl.gov

Abstract

Characterizing the loss of a neural network with respect to model parameters, i.e., the *loss landscape*, can provide valuable insights into properties of that model. Various methods for visualizing loss landscapes have been proposed, but less emphasis has been placed on quantifying and extracting actionable and reproducible insights from these complex representations. Inspired by powerful tools from topological data analysis (TDA) for summarizing the structure of high-dimensional data, here we characterize the underlying *shape* (or topology) of loss landscapes, quantifying the topology to reveal new insights about neural networks. To relate our findings to the machine learning (ML) literature, we compute simple performance metrics (e.g., accuracy, error), and we characterize the local structure of loss landscapes using Hessian-based metrics (e.g., largest eigenvalue, trace, eigenvalue spectral density). Following this approach, we study established models from image pattern recognition (e.g., ResNets) and scientific ML (e.g., physics-informed neural networks), and we show how quantifying the shape of loss landscapes can provide new insights into model performance and learning dynamics.

1 Introduction

Given the important role that the loss function plays during learning, examining it with respect to a neural network's weights—by visualizing the so-called *loss landscape*—can provide valuable insights into both network architecture and machine learning (ML) dynamics [Martin and Mahoney, 2021, Martin et al., 2021, Yang et al., 2022b, 2021, Zhou et al., 2023]. Indeed, the loss landscape has been essential for understanding certain aspects of deep learning, including, but not limited to, test accuracy, robustness of transfer learning [Djolonga et al., 2021], robustness to out-of-distribution

[*] Equal contribution.

Workshop on Scientific Methods for Understanding Deep Learning, NeurIPS 2024.

detection [Yang et al., 2022a], robustness to adversarial attack [Kurakin et al., 2016], and generalizability [Cha et al., 2021]. There are two popular approaches to generating the loss landscape for a given neural network model. Initial efforts to visualize the loss landscape relied on sampling random orthogonal vectors and projecting weights onto the plane spanned by these random vectors [Goodfellow et al., 2014, Li et al., 2018]. More recently, Yao et al. [2020] proposed using directions based on the Hessian, wherein the first two most important Hessian eigenvectors are used to capture more meaningful changes in the loss function. In both approaches, a neural network's parameters are perturbed along each direction, and the loss is re-evaluated at each of these positions.

While both approaches have provided valuable insights, *loss landscape visualization* (no matter which method was used) is often limited to just that—visualization. In other words, loss landscapes, once created, are often simply visually explored or qualitatively compared. It is less clear how to meaningfully measure or quantitatively relate these landscapes to features of the model's underlying architecture or to properties inherent to the learning process. Indeed, examining and quantifying a loss landscape—which is inherently high-dimensional, with as many dimensions as the number of parameters in the model—is challenging to do, especially when using two-dimensional views and qualitative observations alone.

To provide a more quantitative approach to understanding and using loss landscapes, here we show how topological data analysis (TDA) can be used to quantify and extract (quantitative) insights based on the topology (or shape) of those landscapes. We first compute loss landscapes using either random projections or Hessian-based directions and explore four different representations, including one image data representation (where the loss is stored as pixels) and three unstructured grid representations (where the loss is stored on the vertices of a graph). We then apply two methods from TDA, namely, the merge tree [Carr et al., 2003, Heine et al., 2016] and persistence diagram [Edelsbrunner and Harer, 2008], to quantify and compare different loss landscapes. We quantify these structures by measuring the number of saddle points and average persistence, respectively, and we compare our results with state-of-the-art methods for evaluating model performance, as well as with more recent methods for evaluating the local geometry of loss landscapes based on the Hessian.

2 Background on TDA

Topological data analysis (TDA) aims to reveal the global underlying structure of data. It is particularly useful for studying high-dimensional data or functions, where direct visualization is impossible. Here, we leverage ideas and algorithms from TDA to study the structure of the loss function, i.e., the so-called loss landscape. In the context of a loss function, we are interested in the number of minima (i.e., unique sets of parameters for which the loss is locally minimized) and how "prominent" they are (i.e., measuring how many other sets of neighboring parameters have a higher loss than the parameter sets that locally minimize the loss function). Such information can be obtained from the merge tree and persistence diagram (i.e., captured by the 0-dimensional persistent homology).

A *merge tree* [Carr et al., 2003, Heine et al., 2016] tracks connected components of sub-level sets $L^{-}(v) = \{x \in D; x \leq v\}$ as a threshold, v, is increased. Note, the merge tree can track either sub-level or super-level sets, but here we are interested in characterizing loss functions and their minima, so we focus on sub-level sets. In this case, as v increases, new connected components form at local minima and later merge with neighboring connected components (other local minima) at saddles. The merge tree encodes these changes in the loss landscape as nodes in a tree-like structure, where local minima are represented by degree-one nodes (connected to other local minima through a single saddle point), and the saddle points connecting different minima are represented by degree-three nodes (each connecting two local minima and one other saddle point).

A *persistence diagram* represents features (i.e., branches in the merge tree) as points in a two-dimensional plane. The horizontal axis corresponds to the birth of each feature—which is the value of the minimum at which it first appears. The vertical axis corresponds to the death of each feature—which is the value of the saddle where it merges into a more persistent feature. The distance between a point and the diagonal line $y = x$ encodes the persistence of the feature—which is a measure of how long the feature lasts in the filtration (i.e., as the threshold, v, is increased). Such a metric captures information about the ruggedness of the landscape, and for example, the depth of local valleys and the height of the barriers between them. Here, we quantify the average persistence by computing the distance between each persistence pair and the diagonal and then taking the average.

3 Empirical Results

In this section, we provide a summary of our main empirical results. We study established models from image pattern recognition (e.g., ResNets) and scientific ML (e.g., physics-informed neural networks). We first show how removing residual connections from ResNet changes the shape of the loss landscape. We then show how the loss landscape changes for a physics-informed neural network as a physical parameter is varied and optimization begins to fail. In both experiments, we quantify the merge tree and persistence diagrams, and we relate our results to model performance and loss landscape metrics like the top Hessian eigenvalue and the Hessian trace.

3.1 Image Pattern Recognition

In our first experiment, we explore image pattern recognition using ResNet-20 trained on CIFAR-10 [Yao et al., 2020]. Specifically, we look at (and quantify) loss landscapes before and after adding residual connections. Recent work shows that the residual connections are related to the "smoothness" of the loss landscape [Li et al., 2018, Yao et al., 2020]. Here, we aim to verify this and further characterize how the residual connections in ResNet-20 change the underlying loss landscape. Note, we removed the residual connections from ResNet-20 before training. The accuracy of ResNet-20 without residual connections (90% average accuracy across four random seeds) was slightly lower than ResNet-20 with residual connections (92% average accuracy across four random seeds).

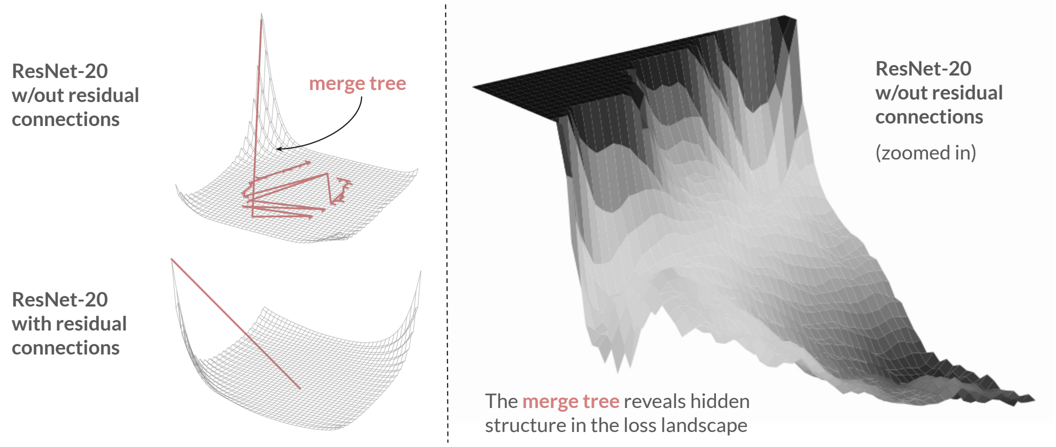

Figure 1: Visualizing loss landscapes for ResNet-20. Compare with Yao et al. [2020].

In Fig. 1, we show loss landscapes for ResNet without and with residual connections. We observe a much more complicated structure in the loss landscape for ResNet-20 without residual connections, revealed by the more complicated branched structure of the merge tree. Interestingly, this complicated structure is not immediately visible in the loss landscape itself. On the right, we show the same landscape for ResNet-20 without residual connections, after clipping outlier values to "zoom" into the smaller scale structure. The misleading visualization on the left highlights a unique benefit of our approach. That is, the merge tree can capture interesting structure across different scales by default, requiring less manual fine-tuning of visualization parameters. Comparatively, we observe a much smoother loss landscape for ResNet-20 with the residual connections, confirmed by the simpler structure of the merge tree (i.e., a single minima and no saddle points). Interestingly, these observations agree with previous findings that adding residual connections to ResNet results in a "smoother" loss landscape, thereby improving generalization [Li et al., 2018].

In Fig. 5 (in the appendix), we further verify these observations numerically by showing how our TDA-based metrics relate to ML-based metrics. These plots provide additional insights beyond the qualitative differences in the loss landscapes we observed, revealing that the number of saddle points in the merge tree increases and the average persistence decreases when the residual connections are removed from ResNet-20. Looking across the different columns, we observe an inverse relationship between the number of saddle points in the merge tree and the ML-based metrics, but a direct