



1-Dimensional Topological Invariants to Estimate Loss Surface Non-Convexity

D. Voronkova, Serguei Barannikov, E. Burnaev

► To cite this version:

D. Voronkova, Serguei Barannikov, E. Burnaev. 1-Dimensional Topological Invariants to Estimate Loss Surface Non-Convexity. Доклады Академии Наук / Doklady Mathematics, 2024, 108 (S2), pp.S325-S332. <10.1134/S1064562423701569>. <hal-04728371>

HAL Id: hal-04728371

<https://hal.science/hal-04728371v1>

Submitted on 9 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

1-DIMENSIONAL TOPOLOGICAL INVARIANTS TO ESTIMATE LOSS SURFACE NON-CONVEXITY

© 2023 г. Daria Voronkova^{1,*}, Serguei Barannikov^{1,3}, Evgeny Burnaev^{1,2}

We utilize the framework of topological data analysis to examine the geometry of loss landscape. With the use of topology and Morse theory, we propose to analyse 1-dimensional topological invariants as a measure of loss function non-convexity up to arbitrary re-parametrization. The proposed approach uses optimization of 2-dimensional simplices in network weights space and allows to conduct both qualitative and quantitative evaluation of loss landscape to gain insights into behavior and optimization of neural networks. We provide geometrical interpretation of the topological invariants and describe the algorithm for their computation. We expect that the proposed approach can complement the existing tools for analysis of loss landscape and shed light on unresolved issues in the field of deep learning.

1. INTRODUCTION

Investigation of the properties of loss landscape has become a rapidly developing research direction aimed to better understand the behavior of neural networks. Among other directions, recent works explored the role of architecture design on the structure of loss surface [1], [2], the impact of batch size during training and sharpness of observed local minimum and its generalization ability [3], [4]. However, these works focus on local behavior of loss function in the vicinity of a local minimum. Another line of research explores the property of mode connectivity [5], [6], [7], [8] that is the ability to connect two local minima with a low-error path or subspace. Finally, there is a direction of research that analyses both local and global structure of loss landscape to gain insights about optimization process and characterize successful solutions [9], [10].

Following this tendency, our work explores the loss surface global topological structure through measuring 1-dimensional topological invariants. We hypothesize that topological stability of loss landscape contributes to the generalization ability of a neural network. The proposed approach does not rely on exact parametrization of neural network's weights which serves as an obstacle for some approaches [4]. With the help of topology, it is possible to characterize the global divergence of loss function from being convex even up to an arbitrary continuous reparametrization. We empirically observe that this correlates with neural network model's generalization.

Our contributions are as follows:

- We formulate theoretical foundations for examining loss landscape geometry that are based on topology and Morse theory. As it follows from theoretical studies, topological structure of a given function can be fully described via the set of its topological invariants, the barcodes.
- We provide geometrical interpretation of the 1-dimensional topological invariants and the algorithm for their computation. We elaborate on the practical usage of these invariants and its application to the problems of loss surface analysis.
- We empirically prove the usefulness of the proposed approach with a set of experiments conducted on deep neural networks.
- We release the code for reproducing the results of this paper at <https://github.com/VoronkovaDasha/1-dim-Topological-Invariants>.

2. RELATED WORKS

The loss function and its landscape is the core object in design and optimization of deep neural networks, and there are a lot of attempts to investigate it from various viewpoints. In this section we review several directions that have the most tight relationship to the current work.

¹Skolkovo Institute of Science and Technology, Moscow, Russia

²Artificial Intelligence Research Institute, Moscow, Russia

³CNRS, IMJ, Paris Cité University, France

*E-mail: Darya.Voronkova@skoltech.ru

A straightforward approach to gain insights into training of deep neural networks is via a visualization of loss landscape. 1-D visualizations are often used to explore optimization trajectory [12], sharpness of found solutions [3], existence of high loss barriers between solutions [12], [13], asymmetrical loss directions in the vicinity of local minima [14]. Using 2-D visualizations, the work [1] observes that residual connections prevent loss landscape from transition to chaotic behavior as depth increases. The paper [2] reveals that visual transformers typically have flatter solutions than residual networks through 2-D loss surface visualization. Despite the popularity of loss landscape visualization techniques, it uses a big dimensionality reduction and requires careful interpretation. In particular, the convexity in the loss surface visualized plot does not necessarily imply the convexity of the true high-dimensional loss function [1]. Moreover, though visualization techniques are suitable for qualitative analysis, they do not provide numerical characteristic of loss surface general convexity.

In general case, linear interpolation reveals a high loss barrier between two local minima obtained through gradient-based optimization [12], [5], [6]. However, [5], [6] empirically observe existence of non-linear trainable paths of low loss between two local minima. This observation is often referred to as «mode connectivity». [19] give theoretical explanation of this phenomenon assuming dropout and noise stability of deep neural net. As shown in [15], the property of mode connectivity holds even when the two local minima are obtained through different training procedures. Moreover, [8] experimentally verifies existence of low loss subspaces connecting a set of solutions. The work [6] proposes novel techniques to ensembling of deep neural networks based on the observed property of mode connectivity. Other works [16], [17] explore structure of loss surfaces through search of various patterns under different setups. While these approaches are powerful tools to inspect loss surface, they lack numerical measure quantifying the «badness» of loss surfaces of a deep neural network.

Extending the idea of mode connectivity, [9], [10] argue that solutions of stochastic gradient descent (SGD) are likely to have no barriers in linear interpolation when permutations of DNNs' weights are taken into account. This implies that loss landscape essentially has one basin modulo permutations.

Another line of research relates generalization with sharpness of local minima. Common approaches to measure sharpness [3], [1] have certain drawbacks [4] and can not explain generalization completely. The work [14] reveals that there are many asymmetric directions with different speed of loss increase in the vicinity of a local minimum. Nevertheless, these approaches measure sharpness of each local minimum independently [11], i.e. they do not provide a characteristic for the whole loss landscape.

3. GEOMETRICAL DESCRIPTION AND COMPUTATION

In this section, we describe in geometric terms the topological invariants and the algorithm for their computation in the setup of deep neural networks.

Consider training of an arbitrary neural network f with parameters $\theta \in \Theta \subset \mathbb{R}^d$ on the dataset X with minimization of the empirical risk loss function L . While the loss function L depends both on the parameters θ and data X , in this work we assume that the dataset is fixed and examine only the dependence on the parameters: $L = L(\theta)$. Consider a set of three terminal points $C_0 = \{\theta_{i_1}, \theta_{i_2}, \theta_{i_3} | \theta_{i_k} \in \Theta, \nabla L(\theta_{i_k}) = 0\}$ to which gradient-based optimization processes converge.

For each pair of points $\theta_{i_j}, \theta_{i_k}$ define

$$h_{j,k} = \min_{\substack{\gamma: [0,1] \rightarrow \Theta, \\ \gamma(0)=\theta_{i_j}, \gamma(1)=\theta_{i_k}}} \max_t L(\gamma(t)) \quad (1)$$

Informally, $h_{j,k}$ determines the lowest barrier among all paths connecting two critical points. For the set C_0 , define $h_{C_0}^1 = \max_{j,k} h_{j,k}$.

Let σ denotes the standard 2-dimensional simplex given by equation $t_1 + t_2 + t_3 = 1, t_i \geq 0$, and $e_i, i = 1, 2, 3$, denotes the three vertices of the standard simplex σ .

For the set C_0 define

$$h_{C_0}^2 = \min_{\substack{\zeta: \sigma \rightarrow \Theta \\ \zeta(e_1)=\theta_{i_1}, \zeta(e_2)=\theta_{i_2}, \zeta(e_3)=\theta_{i_3}}} \max L(\zeta(t)) \quad (2)$$

So, $h_{C_0}^2$ denotes the lowest peak of L on 2-simplices, the boundary of which consists of the cycle of the three lowest loss curves connecting the points $\theta_{i_1}, \theta_{i_2}, \theta_{i_3}$ in cyclic order. From topological point of view, points of C_0 constitute a topological feature in the multi-scale sublevel topology of loss function L . This feature, the cycle, appears at level $h_{C_0}^1$ and disappears at level $h_{C_0}^2$. Hence, for the set C_0 the segment $s_{C_0} = [h_{C_0}^1; h_{C_0}^2]$ records the times of 'birth' and 'death' of this topological feature. The length of the segment ('lifespan') determines the importance of the feature: large lifespan indicates important features

while features with small lifespan are similar to noise. When these segments are aggregated over all triplets of minima critical points of function L , one gets an estimate for 1-dimensional barcode of function L :

$$\text{Barcode}_e^1(L) = \sqcup_{C_0} [h_{C_0}^1, h_{C_0}^2] \quad (3)$$

The exact definition of $\text{Barcode}^1(L)$ involves the filtered simplicial complex and is given in section 4. Index 1 indicates that the barcode is computed for topological features obtained as mapping from 2-dimensional standard simplex, whose boundary is a 1-dimensional cycle. Similarly, it is possible to define and compute $\text{Barcode}^0(L)$ whose computation involves only paths between minima critical points as it was shown in [18]. However, in this work we focus on index 1 barcodes.

Barcodes can be used to estimate the topological similarity of two functions. In this work, we focus on divergence of loss function from being convex since gradient-based optimization for a convex function becomes straightforward.

Index 1 barcode of a convex function L_{ideal} is an empty set. The barcodes are invariant under arbitrary continuous reparametrization.

It follows that if L can be represented as a convex function after a reparametrization, then barcodes of L in each dimension greater than zero must be empty. If L is close to a convex up to a reparametrization function, then these barcodes contain only segments of small length, i.e. the corresponding topological features have small lifespans.

Persistence diagrams, which are the equivalent presentations of the barcodes via sets of (birth, death)-points, can be used for qualitative analysis of barcodes: the closer the (birth, death)-points to the diagonal, the smaller their lifespans, the less important the corresponding features are. Quantitative comparison of two functions is possible through the distance between their barcodes. One of the standard choices for distance on the space of barcodes is Bottleneck distance, also known as \mathbb{W}_∞ , Wasserstein- ∞ distance:

$$\mathbb{W}_\infty(D, D_{ideal}) = \inf_{\pi \in \Gamma(D, D_{ideal})} \sup_{a \in D \cup \Delta} \|a - \pi(a)\|, \quad (4)$$

where D is the barcode $\text{Barcode}^1(L)$ of loss function L represented as a cloud of points, D_{ideal} is the barcode $\text{Barcode}^1(L_{ideal})$ of ideal convex loss function L_{ideal} , Δ denotes the diagonal in \mathbb{R}^2 , $\Gamma(D, D_{ideal})$ defines a set of bijections between $D \cup \Delta$ and $D_{ideal} \cup \Delta$. For functions with similar topology, the distance between barcodes is small. It can be shown that if $D = \{(b_i, d_i)\}_{i \geq 1}$ is the set of (birth, death) points, then $\mathbb{W}_\infty(D, D_{ideal}) = \max_i \frac{d_i - b_i}{\sqrt{2}}$.

Computation. Based on the geometrical description, we propose the following procedure to compute index 1 barcodes. Following the [6], we train curves $\phi : [0, 1] \rightarrow \Theta$ between a pair of minima in the form of a polygonal chain:

$$\phi_\theta(t) = \begin{cases} 2(t\theta + (0.5 - t)\theta_1), & 0 \leq t \leq 0.5, \\ 2((t - 0.5)\theta_2 + (1 - t)\theta), & 0.5 \leq t \leq 1, \end{cases} \quad (5)$$

where θ_1, θ_2 are the (fixed) endpoints of the curve, θ are (trainable) parameters of the curve parametrization. We define the similar parametrization for a map $\psi : \sigma \rightarrow \Theta$ on the standard 2-dimensional simplex.

$$\psi_{\bar{\theta}}(t) = \begin{cases} (t_3 - t_2)\theta_1 + 3t_1\theta_{123} + (2t_2 - 2t_1)\theta_{12}, & t_1 \leq t_2 \leq t_3, \\ (t_2 - t_3)\theta_2 + 3t_1\theta_{123} + (2t_3 - 2t_1)\theta_{12}, & t_1 \leq t_3 \leq t_2, \\ (t_1 - t_2)\theta_3 + 3t_3\theta_{123} + (2t_2 - 2t_3)\theta_{23}, & t_3 \leq t_2 \leq t_1, \\ (t_3 - t_1)\theta_1 + 3t_2\theta_{123} + (2t_1 - 2t_2)\theta_{13}, & t_2 \leq t_1 \leq t_3, \\ (t_2 - t_1)\theta_2 + 3t_3\theta_{123} + (2t_1 - 2t_3)\theta_{23}, & t_3 \leq t_1 \leq t_2, \\ (t_1 - t_3)\theta_3 + 3t_2\theta_{123} + (2t_3 - 2t_2)\theta_{13}, & t_2 \leq t_3 \leq t_1, \end{cases} \quad (6)$$

where $\theta_1, \theta_2, \theta_3$ are the (fixed) vertices of the triangle, $\theta_{12}, \theta_{13}, \theta_{23}, \theta_{123}$ are trainable parameters of the triangle parametrization. In our experiments, we perform training in three stages. First, we obtain $\theta_1, \theta_2, \theta_3$ by standard training of a deep neural network. Second, for each pair of minima, we train a connecting path with polygonal parametrization in Equation 5 and obtain $\theta_{12}, \theta_{13}, \theta_{23}$. Finally, for this set of local minima, we train a connecting triangle and obtain θ_{123} . When training a triangle, we initialize $\theta_1, \theta_2, \theta_3$ with weights corresponding to local minima. $\theta_{12}, \theta_{13}, \theta_{23}$ are initialized with weights corresponding to learned connecting curves. Initialization of θ_{123} is linear interpolation of $\theta_1, \theta_2, \theta_3$. $\bar{\theta}$ is a set of trainable and fixed parameters: $\bar{\theta} = \theta \sqcup \theta_f$. In our experiments we set $\theta_1, \theta_2, \theta_3, \theta_{12}, \theta_{13}, \theta_{23}$ to be fixed and θ_{123} to be trainable, i.e. $\theta = \{\theta_{123}\}$ and $\theta_f = \{\theta_1, \theta_2, \theta_3, \theta_{12}, \theta_{13}, \theta_{23}\}$. For update scheme during training a triangle, please refer to Algorithm 1.

Algorithm 1: Optimization of triangle

Input: N - number of iterations, $\bar{\theta} = \theta \sqcup \theta_f$ - a set of trainable and fixed parameters for triangle $\psi_{\bar{\theta}}$ (6), η - learning rate, L - loss function.
 $\theta^0 = \theta$
for $i \leftarrow 1, N$:
 $t_1^i, t_2^i \sim U[0, 1]$
if $t_1^i + t_2^i > 1$ **then**
 $t_1^i, t_2^i = 1 - t_2^i, 1 - t_1^i$
 $t_3^i = 1 - t_1^i - t_2^i$
 $t^i = (t_1^i, t_2^i, t_3^i)$, sample from 2-dimensional standard simplex
 $\theta^i \leftarrow \theta^{i-1} - \eta \text{grad}_{\theta} L(\psi_{\bar{\theta}}(t^i))$
Return: θ^N

To compute then the barcode based on learned triangles, we make a discretization T of 2-dimensional standard simplex σ , consisting of n points. Next, for each learned connecting triangle, we compute the lifespan of the corresponding topological feature with Algorithm 2. Estimated barcode is formed by the union of these segments.

Algorithm 2: Lifespan computation

Input: $\bar{\theta} = \theta \sqcup \theta_f$ - a set of learned parameters, L - loss function, T - discretized 2-dimensional standard simplex.
 $h_{1,2} = \max_{(0,t_2,t_3) \in T} L(\psi_{\bar{\theta}}(t_1, t_2, t_3))$
 $h_{1,3} = \max_{(t_1,0,t_3) \in T} L(\psi_{\bar{\theta}}(t_1, t_2, t_3))$
 $h_{2,3} = \max_{(t_1,t_2,0) \in T} L(\psi_{\bar{\theta}}(t_1, t_2, t_3))$
 $h^1 = \max\{h_{1,2}, h_{1,3}, h_{2,3}\}$ - birth time of topological feature
 $h^2 = \max_{(t_1,t_2,t_3) \in T} L(\psi_{\bar{\theta}}(t_1, t_2, t_3))$ - death time of topological feature
Return: $[h^1, h^2]$

4. LIFETIMES OF DIMENSION r TOPOLOGICAL FEATURES, INDEX r BARCODE VIA GRADIENT DESCENT ON SIMPLICES

Here we describe an algorithm for barcodes for topological features of any dimension $r \geq 1$. These barcodes quantify the index r critical points "lifetimes". The general definition of the higher barcodes is based on the evolution of higher dimensional topological features (cycles, 2d-voids etc) in the sublevel sets of the loss. The r -th barcode records the appearance-disappearance intervals of r -dimensional cycles.

For the calculation of the higher barcodes one first optimizes $(r+1)$ -simplices in the parameter space, similarly to the optimization of triangles, applying the gradient to a random point from the simplex.

The set of optimized simplices is constructed inductively starting from a set of optimized 0-simplices (minima), then the set of optimized 1-simplices for each pair of minima; then optimized 2-simplices for every triple of sampled minima etc.

After optimizing a set of simplices, we calculate the filtration on this set defined as the maximal value of loss on the given simplex. Next step is the construction of filtered simplicial complex from the optimized simplices. The linear boundary operator ∂_r acts on the linear space formally generated by $(r+1)$ -simplices by sending a simplex to the alternated sum of its r -faces. Next the filtration is used to calculate the appearance and disappearance of topological features, via bringing the set of the boundary linear operators ∂_r to simple form [23, 22, 21]. Then the index r barcode is given by the set of births and deaths of r -dimensional topological features in the filtered simplicial complex.

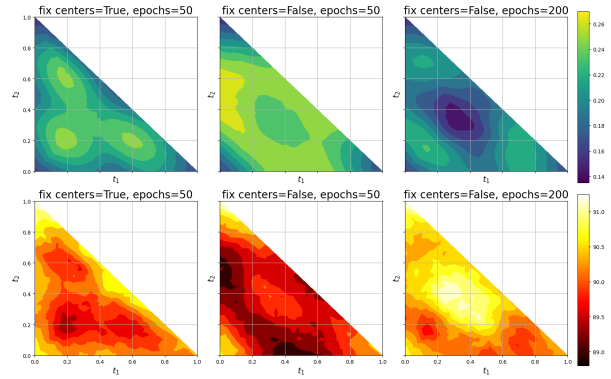


Figure 1: Visualization of the loss function (top) and accuracy (bottom) inside the learned triangle for LeNet-like and FashionMNIST dataset. The title of each subfigure contains the regime of training - trainable or fixed sides' centers - and number of training epochs.

(a) \mathbb{W}_∞ -distance computed for both ResNet-like and VGG-like models.

Type	ResNet-20	ResNet-56 (WideResNet-56-1)	ResNet-110	WideResNet-56-2
with shortcut	0.083	0.061	0.034	0.025
no shortcut (NS)	0.109	0.078	0.0	0.082

The distance \mathbb{W}_∞ for index r barcodes is defined by the same formula 4. Zero distances for all indexes indicate that the loss function, after a change of parametrization in Θ , can be possibly made convex.

5. EXPERIMENTS

In this section we apply the proposed algorithm to explore the topology of loss functions for neural networks.

Training details. We train LeNet-like architecture on FashionMNIST for 50 epochs with batch size 128, weight decay 10^{-3} , learning rate 10^{-2} and learning rate schedule from [6]. To reproduce the experiment setup from [1], similarly, we consider ResNet-like and VGG-like neural networks trained on CIFAR10 dataset. ResNet-like networks are ResNet-20, ResNet-56, ResNet-110 which have respectively 20, 56, 100 layers. VGG-like networks are produced from ResNet-like networks by removing shortcut connections, named with additional suffix 'no short': ResNet-20-NS, ResNet-56-NS, ResNet-110-NS. We also use WideResNet-56-k where $k=1, 2$ means k times more filters per layer. WideResNet-56-1(-NS) is essentially ResNet-56(-NS). To compute barcode, we follow the procedure described in section 3. For training of all intermediate structures, our training procedures are identical to training procedure for minima.

Results. First, we provide experiments on LeNet-like network and FashionMNIST dataset as a proof of concept for the method of learning triangles connecting triples of minima in the network weight space. In Figure 1, we experiment with both regimes of work (fixed and trainable sides' centers) and with different computational budgets for LeNet-like model and FashionMNIST dataset. We find out that the more flexible model leads to better performance but requires bigger computational cost.

Next, we provide experiments with deep neural networks. Similarly to [1], we explore the impact of skip connections when the depth or width of the network increases. The effect of increasing depth on ResNet-like and VGG-like networks is shown in Figures 2, 4a, Table 1a. We can note that loss function of networks with skip connections typically have topological features with smaller lifespans, hence, they are closer to the convex function. Also, we can note that for ResNet110-NS the lifespan of topological feature is zero. In this case, additionally we should consider topological invariants in lower/higher dimensions. For the effect of increasing width on WideResNet-56-1, WideResNet-56-2 and WideResNet-56-1-NS, WideResNet-56-2-NS, see Figures 3, 4b, Table 1a. Loss functions of ResNet-like networks are closer to being convex that can be see from both Figure 2 and Table 1a. In general, increasing depth or width in ResNet-like networks results not only in lowering the \mathbb{W}_∞ distance on the corresponding barcodes, but also in lowering the barriers around the local minima in the loss function landscape.

6. DISCUSSION

In this section we discuss the practical issues and the scope of applicability of the proposed approach.

According to the notion of barcode, it involves computations for each critical point of a loss function. In the case of neural networks, it is hardly possible to reach all critical points during optimization in practice. A standard solution in topological data analysis is to use samples of critical points. Stability of persistence diagrams computed with a point cloud is thoroughly studied in [20]. Moreover, we propose to perform an estimation of barcode only for those points that are reachable by gradient-based optimization, which is in line with practical interest.

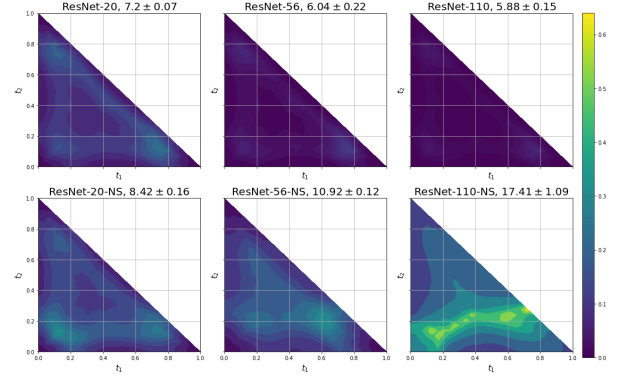


Figure 2: Visualization of the loss function inside the learned triangle for ResNet and ResNet-noshort with different depth. The title of each subfigure contains model specification and test error.

One of the possible applications of the proposed approach is visualization of loss landscape. While [1] provides an approach for visualization through random directions in the parameter space, it has certain drawbacks. At the same time, it could be more informative to understand how the optimization process traverses the surface and whether or not it finds areas with high barriers. In this case, the proposed approach is more suitable as it reveals irreducible obstacles on the way of gradient descent. Moreover, using the proposed \mathbb{W}_∞ distance on the space of barcodes, it is possible to quantify the global non-convexity of the loss landscape. This may be relevant when visual inspection is impossible or controversial.

Moreover, the proposed approach relies on both the parameter space and optimization procedure, hence, it explores the landscape in a more complex way. This view is a more practical since it allows to gain insight in both optimization hyperparameters tuning and architecture design.

7. CONCLUSION

With the help of topology, we provide a novel approach for loss landscape analysis that is based on 1-dimensional topological invariants as measure of loss function parametrization invariant non-convexity. We propose a geometrical description of these features and demonstrate the algorithm for their computation in practice. Also, we verified the proposed approach on a set of deep neural networks. According to our results, in most cases, the proposed approach captures the topology of loss landscape. However, it may further benefit from combining with lower/higher dimensional topological invariants. We discuss several use cases where our approach can be applicable. Hence, we expect it to contribute to the loss landscape exploration and promote topological analysis of loss function.

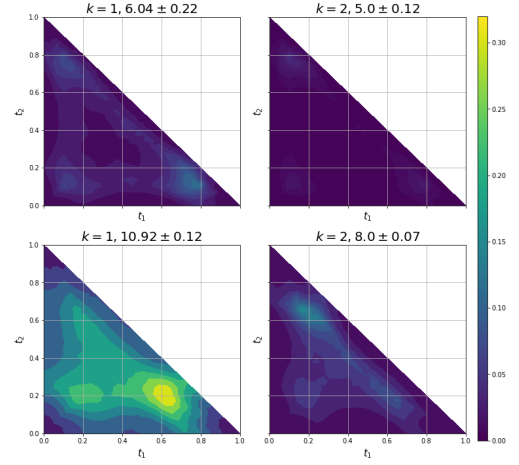
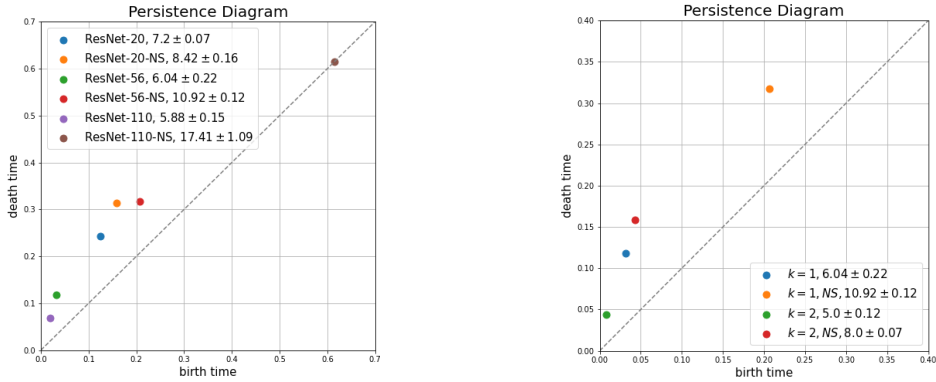


Figure 3: Visualization of the loss function inside the learned triangle for WideResNet-56-k and WideResNet-56-k-noshort with different width. The title of each subfigure contains width specification k and test error.



(a) ResNet and ResNet-noshort with different depth. (b) WideResNet-56-k and WideResNet-56-k-noshort with different width. The title of each subfigure contains model specification and test error.

Figure 4: Persistence diagrams.

REFERENCES

- [1] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, Tom Goldstein. Visualizing the Loss Landscape of Neural Nets. NeurIPS 2018.
- [2] Namuk Park, Songkuk Kim. How Do Vision Transformers Work? ICLR 2022.
- [3] Nitish Shirish Keskar, Dhruv Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, Ping Tak Peter Tang. On Large-Batch Training For Deep Learning: Generalization Gap and Sharp Minima. ICLR 2017.
- [4] Laurent Dinh, Razvan Pascanu, Samy Bengio, Yoshua Bengio. Sharp Minima Can Generalize For Deep Nets. ICML 2017.

- [5] *Felix Draxler, Kambis Veschgini, Manfred Salmhofer, Fred A. Hamprecht.* Essentially No Barriers in Neural Network Energy Landscape. ICML 2018.
- [6] *Timur Garipov, Pavel Izmailov, Dmitrii Podoprikin, Dmitry Vetrov, Andrew Gordon Wilson.* Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. NeurIPS 2018.
- [7] *Gregory W. Benton, Wesley J. Maddox, Sanae Lotfi, Andrew Gordon Wilson.* Loss Surface Simplexes for Mode Connecting Volumes and Fast Ensembling. ICML 2021.
- [8] *Stanislav Fort, Stanislaw Jastrzebski.* Large Scale Structure of Neural Network Loss Landscapes. 2019.
- [9] *Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur.* The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks. ICLR 2022.
- [10] *Samuel K. Ainsworth, Jonathan Hayase, Siddhartha Srinivasa.* Git Re-Basin: Merging Models Modulo Permutation Symmetries. ICLR 2023.
- [11] *Yaoqing Yang, Liam Hodgkinson, Ryan Theisen, Joe Zou, Joseph E. Gonzalez, Kannan Ramchandran, Michael W. Mahoney.* Taxonomizing Local Versus Global Structure in Neural Network Loss Landscapes. NeurIPS 2021.
- [12] *Ian J. Goodfellow, Oriol Vinyals, Andrew M. Saxe.* Qualitatively Characterizing Neural Network Optimization Problems. ICLR 2015.
- [13] *Leslie N. Smith, Nicholay Topin.* Exploring Loss Function Topology with Cyclical Learning Rates. ICLR 2017.
- [14] *Haowei He, Gao Huang, Yang Yuan.* Asymmetric Valleys: Beyond Sharp and Flat Local Minima. 2019.
- [15] *Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, Richard Socher.* Using Mode Connectivity for Loss Landscape Analysis. 2018.
- [16] *Ivan Skorokhodov, Mikhail Burtsev.* Loss Surface Sightseeing by Multi-Point Optimization. NeurIPS 2019.
- [17] *Wojciech Marian Czarnecki, Simon Osindero, Razvan Pascanu, Max Jaderberg.* A Deep Neural Network's Loss Surface Contains Every Low-dimensional Pattern. 2020.
- [18] *Serguei Barannikov, Alexander Korotin, Dmitry Oganessian, Daniil Emtsev, Evgeny Burnaev.* Barcodes as summary of loss function's topology. 2020.
- [19] *Rohith Kuditipudi, Xiang Wang, Holden Lee, Yi Zhang, Zhiyuan Li, Wei Hu, Sanjeev Arora, Rong Ge.* Explaining Landscape Connectivity of Low-cost Solutions for Multilayer Nets. NeurIPS 2019.
- [20] *Chazal, Frédéric and Michel, Bertrand* An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *Frontiers in artificial intelligence*, v. 4, p. 108, 2021.
- [21] *Le Peutrec, D. and Nier, F. and Viterbo, C.* Precise Arrhenius Law for p-forms: The Witten Laplacian and Morse–Barannikov Complex", journal="Annales Henri Poincaré. *Annales Henri Poincaré*, 2013.
- [22] *Zomorodian, Afra J.* Computing and comprehending topology: Persistence and hierarchical Morse complexes (Ph.D.Thesis). 2001.
- [23] *Serguei Barannikov.* Framed Morse complexes and its invariants. *Advances in Soviet Mathematics*, v. 21, 93-116, 1994.