Barcodes as Summary of Loss Function Topology

Serguei Barannikov, A. Korotin, D. Oganesyan, D. Emtsev, E. Burnaev

ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2023, том , с.

МАТЕМАТИКА

BARCODES AS SUMMARY OF LOSS FUNCTION TOPOLOGY

© 2023 г. Serguei Barannikov[1,3,*], Alexander Korotin[1,2], Dmitry Oganesyan[1], Daniil Emtsev[1,4], Evgeny Burnaev[1,2]

We propose to study neural networks' loss surfaces by methods of topological data analysis. We suggest to apply barcodes of Morse complexes to explore topology of loss surfaces. An algorithm for calculations of the loss function's barcodes of local minima is described. We have conducted experiments for calculating barcodes of local minima for benchmark functions and for loss surfaces of small neural networks. Our experiments confirm our two principal observations for neural networks' loss surfaces. First, the barcodes of local minima are located in a small lower part of the range of values of neural networks' loss function. Secondly, increase of the neural network's depth and width lowers the barcodes of local minima. This has some natural implications for the neural network's learning and for its generalization properties.

1. INTRODUCTION

Searching for minima of the loss function is the principal strategy underlying the majority of machine learning algorithms. The graph of the loss function, which is often called **loss surface**, typically has complicated structure [1, 2, 3]: non-convexity, many local minima, saddle points, flat regions. These obstacles harm the exploration of the loss surface.

The searching for minima of modern neural networks is mainly carried out by gradient descent based algorithms. How these algorithms can achieve almost zero loss despite the non-convexity of the loss function remains poorly understood.

The global topological characteristics of the gradient flow trajectories are captured by the Morse complex via decomposing the parameter space into cells of uniform flow [4, 5, 6]. The barcodes of Morse complex constitute the fundamental summary of the topology of the gradient vector field flow [7, 8, 9]. Barcodes give a decomposition of topology change of the loss function sublevel sets into the sum of "birth"-"death" of elementary features.

We argue, see section 5, that, topologically, the "badness" of a given local minimum, harming gradient-based learning algorithms, can be quantified by the "lifetime" length of the minimum's segment in the barcode. For gradient-based learning algorithms, this quantity measures *the obligatory penalty for moving from the given local minimum to a point with lower loss value.*

The calculation of the barcodes for various specific functions constitutes the essence of the topological data analysis. Currently available software packages for the calculation of barcodes of functions, also called "sublevel persistence", are GUDHI, Dionysus, PHAT. They are based on the general algorithm requiring construction of simplicial complex and having $O(N^3)$ worst time complexity in the number of points for computation of the lowest degree barcode. These packages can currently handle calculations of barcodes for functions defined on a grid, of up to 10^6 points, and in dimensions up to six. Thus, all current packages experience the scalability issues.

We describe an algorithm for the computation of lowest degree barcodes for functions in arbitrary dimensions. In contrast to the mentioned grid-based methods, our algorithm works with functions defined on arbitrarily sampled **point clouds**. Point cloud based methods are known to work better than grid-based methods in optimization-related problems[10]. To compute the lowest degree barcodes we use the fact that their definition can be reformulated in geometrical terms, see definition 1 in section 2. Most currently available software packages are based on the more algebraic approach as in definition 2 from section 2. The principal

[1] *Skolkovo Institute of Science and Technology, Moscow, Russia*

[2] *Artificial Intelligence Research Institute, Moscow, Russia*

[3] *CNRS, IMJ, Paris Cité University, France*

[4] *ETH Zurich, Switzerland*

[*] *E-mail: s.barannikov@skoltech.ru*

part of our algorithm has worst time complexity $O(N \log N)$ and it was tested in dimensions up to 15 and with the number of points of up to 10^9.

The proposed methodology describes properties of the loss surface of neural networks via topological features of local minima. We emphasize that the value of the loss at a minimum is only half of its topological characteristic from the barcode. The other half can be described as the value of loss function at the 1-saddle, which is naturally associated with each local minimum, see section 2. The 1-saddle q associated with the minimum p is the point where the connected component of the sublevel set $\Theta_{f\leqslant c} = \{\theta \in \Theta \mid f(\theta) \leqslant c\}$ containing p merges with another connected component of the sublevel set containing a *lower* minimum. This correspondence between local minima and 1-saddles, that kill a connected component of $\Theta_{f\leqslant c}$, is one-to-one.

The segment $[f(p), f(q)]$, where q is the 1-saddle associated with p, is the invariant which the barcode associates with the minimum p. The difference $f(q) - f(p)$ is the topological invariant of the minimum, quantifying its badness for gradient-based learning algorithms. The set of all such segments for all minima is the lowest degree barcode of f.

The main contributions of the paper are as follows:

Applying the minima barcodes and the correspondence between minima and $1-$saddles to exploration of loss surfaces. For each local minimum p there is canonically defined 1-saddle q (see Section 2). The set of all segments $[f(p), f(q)]$, where p is a local minimum and q is the corresponding 1-saddle of f, is a robust topological invariant of loss function. It is invariant in particular under the action of homeomorphisms of Θ. This set of segments is a part of the full barcode. The full barcode gives a concise summary of the topology of the loss function and of the global structure of its gradient flow.

Algorithm for calculations of the minima barcodes. We describe and analyze an algorithm for calculation of the barcodes of minima. The algorithm takes as an input a randomly sampled or a specifically chosen set of points and the loss function's values on this set. The algorithm then employs the HNSW procedure to calculate the graph of neighbors. The next step is the computation of the barcode of minima for the function defined on the graph. The local minima give birth to clusters of points in sublevel sets. Then algorithm works by looking at neighbors of each sampled point with lower values of the function and deciding if this point belongs to an existing cluster, gives birth to a new cluster (minimum), or unifies two or more clusters (1-saddle). The second part of the algorithm working with a function on a graph is similar to a particular $0-$dimensional case of the general algorithm described in [7].

Experiments confirming observations on behavior of neural networks loss functions barcodes. We calculate the barcodes of minima for small fully-connected neural networks of up to three hidden layers and verify that all segments of minima's barcode belong to a small lower part of the total range of loss function's values and that with the increase in the neural network depth the minima's barcodes descend lower.

The usefulness of our approach and algorithms is not limited to optimization problems. Our algorithm permits the fast computation of the persistence barcodes of many functions which were not accessible until now. These sublevel persistence barcodes have been successfully applied in different disciplines: cognitive science [11], cosmology [12] to name a few, see e.g. [13] and references therein.

Our framework also has applications in chemistry and material science where 1-saddle points on potential energy landscapes correspond to transition states and minima are stable states corresponding to different materials or protein foldings, see e.g. [14, 15].

The article is structured as follows. We begin with two definitions of barcodes of local minima in section 2. Our algorithm for the calculation of barcodes is described in section 3. In section 4 we apply our algorithm to calculate barcodes of benchmark functions. We prove the convergence of the algorithm and demonstrate it empirically in subsection 4. In section 5 we calculate barcodes of the loss functions of small neural networks and describe our principal observations.

2. TOPOLOGY OF LOSS SURFACES VIA BARCODES

Barcodes give a concise summary of topological features of functions as decomposition of change of topology of function's sublevel sets into the finite sum of "birth"–"death" of elementary features. We propose to apply these invariants as a tool for exploring the topology of loss surfaces.

We describe in this section two definitions of the barcodes of minima for piecewise-smooth continuous functions. In this work we concentrate on the part of barcodes, describing the "birth"–"death" phenomena of connected components of the loss function's sublevel sets. The approach from this section works similarly in the context of "almost minima", i.e. for the critical points (manifolds) of small nonzero indexes. Such points are often the terminal points of optimization algorithms in extremely high dimensional parameter spaces of deep neural networks [2].

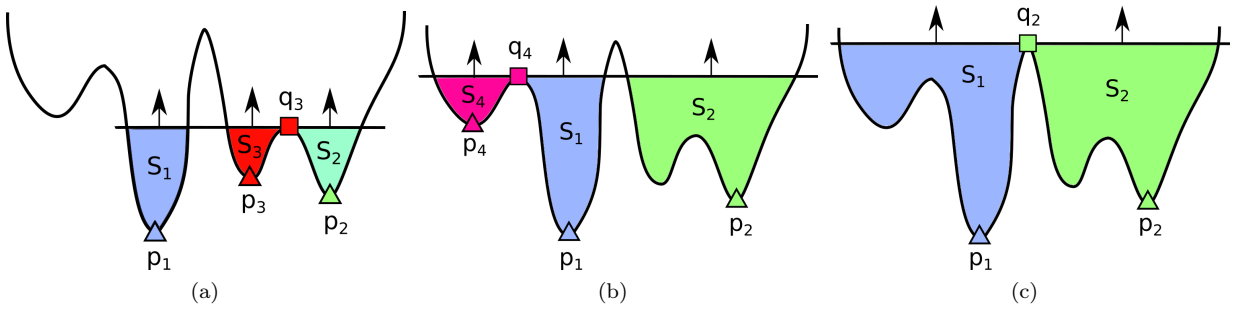
First definition: merging with connected component of a lower minimum.

Let f be a piecewise-smooth continuous function. The values of parameter c at which the topology of sublevel sets

$$\Theta_{f\leqslant c} = \{\theta \in \Theta \mid f(\theta) \leqslant c\}$$

changes are critical values of f.

Let p be one of the minima of f. When c increases from $f(p) - \epsilon$ to $f(p) + \epsilon$, a new connected component of the set $\Theta_{f\leqslant c}$ is born. To illustrate the process, we provide an example in Figure 1, where the connected components S_1, S_2, S_3, S_4 of sublevel set are born at the minima p_1, p_2, p_3, p_4 correspondingly.



(a) (b) (c)

Figure 1: Merging of connected components of sublevel sets at 1-saddles. (a)"Death" of the connected component S_3, the connected component S_3 of sublevel set merges with connected component S_2 at 1-saddle q_3, 1-saddle q_3 is associated with the minimum p_3. (b)"Death" of the connected component S_4, the connected component S_4 of sublevel set merges with connected component S_1 at 1-saddle q_4, 1-saddle q_4 is associated with the minimum p_4. (c)"Death" of the connected component S_2, the connected component S_2 of sublevel set merges with connected component S_1 at 1-saddle q_2, 1-saddle q_2 is associated with the minimum p_2. Note that the 1-saddle q_2 is associated with the minimum p_2 which is separated by another minimum from the green saddle.

Connected components of sublevel sets merge at 1-saddle critical points.

A point q is a *1-saddle* critical point if the intersection of the set $\Theta_{f<f(q)}$ with any small neighborhood of q has more than one connected component. Without loss of generality, we may assume in this section that the 1−saddle points of f are generic so that these intersections have no more than two connected components. Otherwise, one can always add a small perturbation to f or consider the 1−saddles with multiplicity given by the number of these components minus one.

Let p be a minimum, which is not global. When c is increased sufficiently, the connected component of $\Theta_{f\leqslant c}$ born at p merges with some other connected component. Then this unified connected component may merge again with another one. After each merging, the minimum of the restriction of f to the unified connected component is the smallest of the two minima of restriction of f to each of the two connected components before merging. In other words, the connected component with lower minimum "swallows" at the merging point the connected component with the higher minimum, see fig 1. Let q be the merging point where the connected component with minimum p is swallowed by a connected component whose minimum is *lower*. Note that the intersection of the set $\Theta_{f<f(q)}$ with any small neighborhood of q has at least two connected components.

Definition 1. The merging point q, where the connected component with minimum p is swallowed by connected component with a lower minimum, is the 1-saddle associated naturally with the minimum p. The segment $[f(p), f(q)]$ is the invariant associated in barcode with minimum p.

Note that the two connected components of the intersection of a small neighborhood of such q with $\Theta_{f<f(q)}$ belong to two different connected components of the whole set $\Theta_{f<f(q)}$. The 1-saddles of this type are called "+" ("plus") or "death" type.

Proposition 1. *The described correspondence between local minima and 1-saddles of this type is one-to-one.*

Proof. The correspondence in the opposite direction can be described as follows. Let q be a 1-saddle point of such type that the two branches of the set $\Theta_{f<f(q)}$ near q are not connected in the whole set $\Theta_{f<f(q)}$. One of the connected components of the sublevel set $\Theta_{f\leqslant c}$ splits into two when c decreases from $f(q) + \epsilon$ to $f(q) - \epsilon$. Let p_1 and p_2 be the two minima of the restriction of f to each of these two connected components. Let $p_1 > p_2$ be the highest of the two minima. The 1-saddle q is associated by the Definition 1 with the local

minimum p_1, since the two connected components merge at q.

$$p_1 \notin \left\{ \arg \min_{\substack{x \in \text{connected component} \\ \text{of } \Theta_{f \leqslant f(q)+\epsilon}}} f(x) \right\}$$

Notice that p_1 did not appear as a minimum of f on one of connected components of $\Theta_{f \leqslant f(q)+\epsilon}$. In other words the minimum p_1, that corresponds to q by definition 1, is the new minimum appearing in the set of minima of f on connected components of $\Theta_{f \leqslant c}$ when c decreases from $f(q)+\epsilon$ to $f(q)-\epsilon$. □

It is important for application in section 5 that 1-saddles associated with minima are described in the following way.

Proposition 2. *Consider various paths γ starting from local minimum p and going to a lower minimum. Let $m_\gamma \in \Theta$ be the maximum of the restriction of f to such path γ. Then 1-saddle q corresponding to the local minimum p in the barcode is the minimum over the set of all such paths γ of the maxima m_γ:*

$$q = \arg \left[\min_{\substack{\gamma:[0,1]\to\Theta \\ \gamma(0)=p,\, f(\gamma(1))<f(p)}} \max_t f\big(\gamma(t)\big) \right] \tag{1}$$

Proof. Let the connected component of local minimum p merge with the connected component of a lower minimum at the point q. Then there is a path in the sublevel set $\Theta_{f \leqslant f(q)}$ passing through q and connecting p with the lower minimum. The point q is the maximum of the restriction of f to such a path. And there is no path in sublevel sets $\Theta_{f \leqslant c}$, $c < f(q)$, that connects p with a lower minimum. Therefore there is no path with $f(m_\gamma) < f(q)$ and q satisfies equation (1). □

Second definition: invariants of filtered complexes.

We recall here the definition for the full barcode [7, 8]. Although our algorithm from section 3 is based on definition 1, we need definition 2 below to put things into the general framework in which these invariants constitute the full description of the loss function topology and of the global behaviour of the gradient flow.

Families of gradient flow trajectories emanating from the same singular point decompose the domain Θ into cells on which the gradient flow behaves uniformly[6].

Chain complex is the algebraic counterpart of an intuitive idea representing complicated geometric objects as a decomposition into simple pieces. It converts such a decomposition into a collection of vector spaces and linear maps.

Recall that a chain complex (C_*, ∂_*) is a sequence of finite-dimensional vector spaces C_j (spaces of "$j-$chains") and linear operators ("differentials")

$$\to C_{j+1} \overset{\partial_{j+1}}{\to} C_j \overset{\partial_j}{\to} C_{j-1} \to \ldots \to C_0,$$

which satisfy $\partial_j \circ \partial_{j+1} = 0$. The $j-$th homology of the chain complex (C_*, ∂_*) is the quotient of vector spaces $H_j = \ker(\partial_j) / \mathrm{im}(\partial_{j+1})$.

The decomposition of complicated geometric object into simple pieces is often done in certain consecutive order, in that case its algebraic counterpart is $\mathbb{R}-$filtered chain complex. A subcomplex $(C'_*, \partial'_*) \subseteq (C_*, \partial_*)$ is a sequence of subspaces $C'_j \subseteq C_j$ equipped with compatible differentials $\partial'_j = \partial_j \mid_{C'_j}$. A chain complex C_* is called $\mathbb{R}-$filtered if C_* is equipped with an increasing sequence of subcomplexes ($\mathbb{R}-$filtration): $F_{s_1}C_* \subset F_{s_2}C_* \subset \ldots \subset F_{s_{max}}C_* = C_*$, indexed by a finite set of real numbers $s_1 < s_2 < \ldots < s_{\max}$.

Theorem 3. *([7], Section 2) Any $\mathbb{R}-$filtered chain complex C_* can be brought to canonical form, a canonically defined direct sum of $\mathbb{R}-$filtered complexes of two types: one-dimensional complexes with trivial differential $\partial_j(e_i) = 0$ and two-dimensional complexes with trivial homology $\partial_j(e_{i_2}) = e_{i_1}$, by a linear transformation preserving the $\mathbb{R}-$filtration. The resulting direct sum of such simple $\mathbb{R}-$filtered complexes is unique.*

Definition 2. The full barcode is a visualization of the decomposition of an $\mathbb{R}-$filtered complex according to the theorem 3. Each filtered 2-dimensional complex with trivial homology $\partial_j(e_{i_2}) = e_{i_1}$, $\langle e_{i_1} \rangle = F_{\leqslant s_1}$,$\langle e_{i_1}, e_{i_2} \rangle = F_{\leqslant s_2}$ describes a single topological feature in dimension j which is "born" at s_1 and which "dies" at s_2. It is represented by segment $[s_1, s_2]$ in the degree-j barcode. And each filtered one-dimensional complex with trivial differential, $\partial_j e_i = 0$, $\langle e_i \rangle = F_{\leqslant r}$ describes a topological feature in dimension j which is "born" at r and never "dies". It is represented by the half-line $[r, +\infty[$ in the degree-j barcode.

The proof of Theorem 3 is given in [7], Section 2, see also [8, 9]. We describe it in Appendix B for the reader's convenience. Essentially, one can bring an $\mathbb{R}-$filtered complex to the required direct sum of simple $\mathbb{R}-$filtered complexes by induction, starting from the lowest basis elements of degree one, in such a way

that the manipulation of degree j basis elements does not destroy the obtained decomposition into simple $\mathbb{R}-$filtered complexes in degree $j-1$ and in lower filtration pieces in degree j.

Let $f:\Theta\to\mathbb{R}$ be a smooth, or more generally, piece-wise smooth continuous function such that the sublevel sets $\Theta_{f\leqslant c}=\{\theta\in\Theta\mid f(\theta)\leqslant c\}$ are compact.

There are different filtered chain complexes computing the homology of the topological spaces $\Theta_{f\leqslant c}$, the Čech complexes, the simplicial complexes, or the CW-complexes. Without loss of generality, the piece-wise smooth function f can be assumed smooth, otherwise one can always replace f by a smooth approximation. By adding a small perturbation we can also assume that critical points of f are non-degenerate.

One filtered chain complex naturally associated with such function f, with subcomplexes F_sC_* computing homology of sublevel sets $\Theta_{f\leqslant s}$, is the Morse complex, see appendix A and [4, 5] for more details. The Morse complex is defined as follows. The basis elements in the $k-$vector spaces C_j are in one-to-one correspondence with the critical points of f of index j equipped with a choice of orientation. The orientation here is a choice of the orientation of j-dimensional subspace of tangent space at the critical point, on which the Hessian is negative-definite. The matrix of ∂_j consists of the numbers of gradient trajectories, counted with signs, between the index j and the index $(j-1)$ critical points. The natural filtration on Morse complex is given by the values of critical points. Basis elements in F_sC_* correspond to critical points with critical value s and lower.

Proposition 4. *Let p be a minimum, which is not global. Then p represents trivial homology class in Morse complex computing the homology of $\Theta_{f\leqslant c}$ for big enough c , i.e. p is a lower basis element of one of the two-dimensional complexes with trivial homology in the canonical form. Thus p is coupled with a 1-index saddle q. This is the 1-saddle from definition 1, i.e. q is the 1-saddle at which the sublevel set connected component corresponding to p is swallowed by a connected component with lower minimum. The segment $[f(p),f(q)]$ then corresponds to the minimum p by definition 1.*

Proof. Let q be the 1-saddle at which the connected component corresponding to p is swallowed by a connected component with a lower minimum r. The homology $H_0(\Theta_{f\leqslant f(q)-\epsilon})$ are generated linearly by classes of connected components of $\Theta_{f\leqslant f(q)-\epsilon}$. They correspond to the generators of the Morse complex given by minima of restriction of f to each connected component. In the Morse complex computing homology of $\Theta_{f\leqslant f(q)+\epsilon}$, the generator corresponding to the local minimum p equals to the boundary of the generator corresponding to the 1-saddle q plus the generator corresponding to the lower minimum r plus perhaps boundaries of lower than q saddles. Therefore q is coupled with p in the canonical form. □

Similar results hold for other types of filtered complexes representing sublevel homology, like cubical or CW-complexes, for continuous piece-wise smooth functions (see e.g. [16, 17]).

The total number of different topological features in sublevel sets $\Theta_{f\leqslant c}$ of the loss function can be read immediately from the barcode. Namely, the number of intersections of horizontal line at level c with segments in the index j barcode gives the number of independent topological features of dimension j in $\Theta_{f\leqslant c}$, see examples in section 4.

3. AN ALGORITHM FOR CALCULATION OF MINIMA BARCODES

In this section, we describe an algorithm for the calculation of the barcodes of local minima. The algorithm uses definition 1 of barcodes from Section 2 that is based on the evolution on the connected components of the sublevel sets.

To analyse the surface of the given function $f:\Theta\to\mathbb{R}$, we first build its approximation by finite graph-based construction. To do this, we consider a randomly sampled subset of points $\{\theta_1,\dots,\theta_N\}\in\Theta$ and construct a graph with these points as vertices. We connect vertices with an edge if the points are close. Thus, for every vertex θ_n, by comparing $f(\theta_n)$ with $f(\theta_{n'})$ for neighbors $\theta_{n'}$ of θ_n, we are able to understand the local topology near the point θ_n. At the same time, connected components of sublevel sets $\Theta_{f\leqslant c}$ correspond, see proposition 5 below, to connected components of the subgraph $\Xi_{f\leqslant c}$ of points θ_n, such that $f(\theta_n)\leqslant c$.

Two technical details here are the choice of points θ_n and the definition of closeness, i.e. when to connect points by an edge. In our experiments, we sample points uniformly from some rectangular box of interest. To add edges, we compute the oriented k-nearest neighbor graph on the given points, then drop the orientation of edges and check that the distance between neighbors does not exceed $c(D)N^{-\frac{1}{D}}$, where D is the dimension of f's input. We use $k=2D$ in our experiments

We describe now the part of the algorithm that computes barcodes of a function from its graph-based approximation described above. The key idea is to monitor the evolution of the connected components of the sublevel sets of the graph $\Xi_{f\leqslant c}=\{\theta_n\mid f(\theta_n)\leqslant c\}$ for increasing c.

Algorithm 1: Barcodes of minima computation for function on a graph.

Input : Undirected graph $G = (V, E)$; function f on graph vertices.
Output : Barcodes: a list of "birth"-"death" pairs.
$S \leftarrow \{\}$;
for $\theta \in V$ *in increasing order of* $f(\theta)$ **do**
  $S' \leftarrow \{s \in S \mid \exists \theta' \in s$ such that $(\theta, \theta') \in E$ and $f(\theta) > f(\theta')\}$;
  if $S' = \emptyset$ **then**
    $S \leftarrow S \sqcup \{\{\theta\}\}$;
  else
    $f^* \leftarrow \min f(\theta')$ for $\theta' \in \bigsqcup_{s \in S'} s$;
    for $s \in S'$ **do**
      $f^s \leftarrow \min f(\theta')$ for $\theta' \in s$;
      if $f^s \neq f^*$ **then**
        Barcodes $\leftarrow$ Barcodes $\sqcup \{(f^s, f(\theta))\}$;
    end
    $s_{\text{new}} \leftarrow \big(\bigsqcup_{s \in S'} s\big) \sqcup \{\theta\}$;
    $S \leftarrow (S \setminus S') \sqcup \{s_{\text{new}}\}$;
  end
end
for $s \in S$ **do**
  $f^s \leftarrow \min f(\theta')$ for $\theta' \in s$;
  Barcodes $\leftarrow$ Barcodes $\sqcup \{(f^s, \infty)\}$;
end
return Barcodes

For simplicity we assume that points θ are ordered w.r.t. the value of function f, i.e. for $n < n'$ we have $f(\theta_n) < f(\theta_{n'})$. In this case we are interested in the evolution of connected components throughout the process of sequential adding of vertices $\theta_1, \theta_2, \ldots, \theta_N$ to graph, starting from an empty graph. We denote the subgraph on vertices $\theta_1, \ldots, \theta_n$ by Ξ_n. When we add new vertex θ_{n+1} to Ξ_n, there are three possibilities for connected components to evolve:

1. Vertex θ_{n+1} has zero degree in Ξ_{n+1}. This means that θ_{n+1} is a local minimum of f and it forms a new connected component in the sublevel set.
2. All the neighbors of θ_{n+1} in Ξ_{n+1} belong to one connected component in Ξ_n.
3. All the neighbors of θ_{n+1} in Ξ_{n+1} belong to $K \geqslant 2$ connected components $s_1, s_2, \ldots, s_K$ in Ξ_n. Thus, all these components will form a single connected component in Ξ_{n+1}.

In the third case, according to definition 1 of section 2, the point θ_{n+1} is a discrete 1-saddle point. Thus, one of the components s_k swallows all the rest. This is the component which has the lowest minimal value. For other components this gives their barcodes: for $s_i, i \neq k$ the birth-death pair is $\big[\min_{\theta \in s_i} f(\theta); f(\theta_{n+1})\big]$. We summarize the procedure in the algorithm 1.

In the practical implementation of the algorithm, we precompute the values of function f at the vertices of G. Besides that, we use the disjoint set data structure to store and join connected components during the process. We also keep and update the global minima in each component. We did not include these tricks into the algorithm's pseudo-code in order to keep it simple.

Remark 1. *Given a function on a graph one can construct an* $\mathbb{R}-$*filtered chain complex* $C_1 \xrightarrow{\partial_1} C_0$ *as follows. The basis of the vector space of* $0-$*chains* C_0 *is given by the set of graph's vertices and the basis of the space of* $1-$*chains* C_1 *is given by the set of graph's edges, for simplicity we consider the vector spaces over the field* $\mathtt{k} = \{0, 1\}$*. Then the differential of an edge is the sum of its two ends. The filtration is defined by the values of the function. The filtration for an edge is given by the maximum of the function's values on the two ends of the edge. For this particular chain complex the general algorithm from*[7]*, that calculates the barcodes of* $\mathbb{R}-$*filtered complexes, simplifies and its* $0-$*dimensional part essentially gives the described algorithm for barcodes of minima for function on a graph.*

The resulting complexity of the algorithm's principal part is $O(N \log N)$ in the number of points. Here it is important to note that the procedure of graph creation may be itself time-consuming. In our case, the most time consuming operation is nearest neighbor search. In our code, we used efficient HNSW Algorithm for

approximate NN search by [18]. Also since we only take neighbors lying no further than fixed small distance $r = O(N^{-\frac{1}{D}})$, one can use the following simple strategy as well. First, distribute points from the sample over boxes of fixed grid with edges of size r. Then, in order to determine the neighbors, check distances from each point to the points lying only in the same box and in the neighboring boxes of the grid.

4. BARCODES OF BENCHMARK FUNCTIONS AND CONVERGENCE OF THE ALGORITHM

In this section we apply our algorithm to describing the topology of test functions. In subsection 4 we apply the algorithm to visual examples and in subsection 4 we check the convergence on benchmark functions. In section 5 we apply our algorithm to analysis of loss surfaces of **small neural networks**.

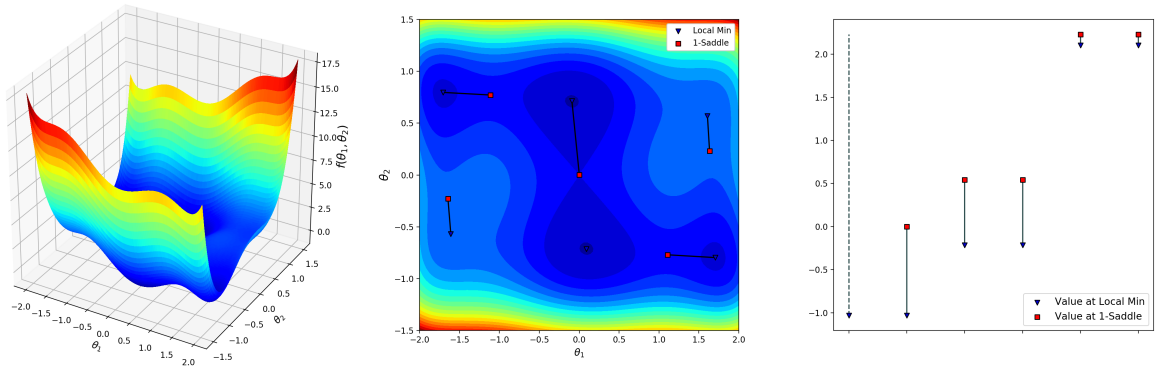
We apply our algorithm to several functions $f : \mathbb{R}^D \to \mathbb{R}$ from **Global Optimization Benchmark** (see e.g. [19]). These functions are designed to fool global optimization algorithms, they are very complex, have many local minima and saddle points even for small dimensions. Thus, the computation of barcodes and minimum-saddle correspondence for these functions is also challenging.

Barcodes of 2D benchmark functions

To begin with, we compute barcodes of several 2-dimensional benchmark functions. We visualise obtained barcodes and minima-saddles correspondence. Next, we conduct the experiments to estimate how the number of points used to compute barcodes influences the quality of the answer. We consider the following test objective functions:

1. **HumpCamel6** function $f : [-2, 2] \times [-1.5, 1.5] \to \mathbb{R}$ with 6 local minima (Figure 2):

$$f(\theta_1, \theta_2) = (4 - 2.1\theta_1^2 + \theta_1^4/3)\theta_1^2 + \theta_1\theta_2 + (-4 + 4\theta_2^2)\theta_2^2$$



(a) Surface plot (3D). (b) Correspondence between minima and 1-saddles (c) Barcode of local minima.

Figure 2: HumpCamel6 function, its minima-saddle correspondence and barcode computed by Algorithm 1.

2. **Langermann** test objective function $f : [0, 10]^2 \to \mathbb{R}$ (Figure 3):

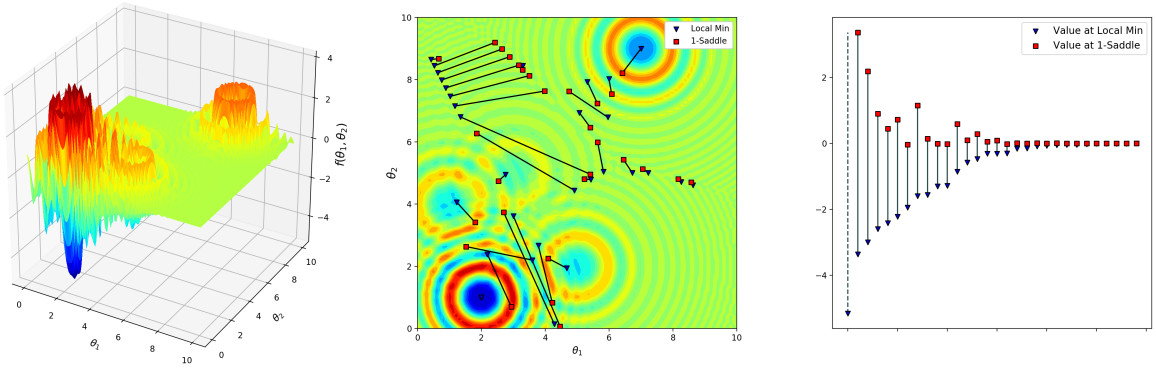
$$f(\theta_1, \theta_2) = -\sum_{i=1}^{5} \frac{c_i \cos\left(\pi\left[(\theta_1 - a_i)^2 + (\theta_2 - b_i)^2\right]\right)}{\exp\left(\frac{1}{\pi}\left[(\theta_1 - a_i)^2 + (\theta_2 - b_i)^2\right]\right)}.$$

   To keep the plots simple, we displayed minimum-saddle correspondence only for Top-30 segments in barcode sorted by the size of the cluster (number of points at the cluster death moment). As wee see in Figure 3b, the correspondence between minimum-saddle is non-trivial. For many minima, the corresponding saddles are rather distant. In particular, this observation illustrates the remark discussed in Figure 1: the corresponding canonical 1-saddle is not necessarily the nearest saddle.

3. **Wavy** test objective function $f : [-\pi, \pi]^2 \to \mathbb{R}$ (Figure 4):

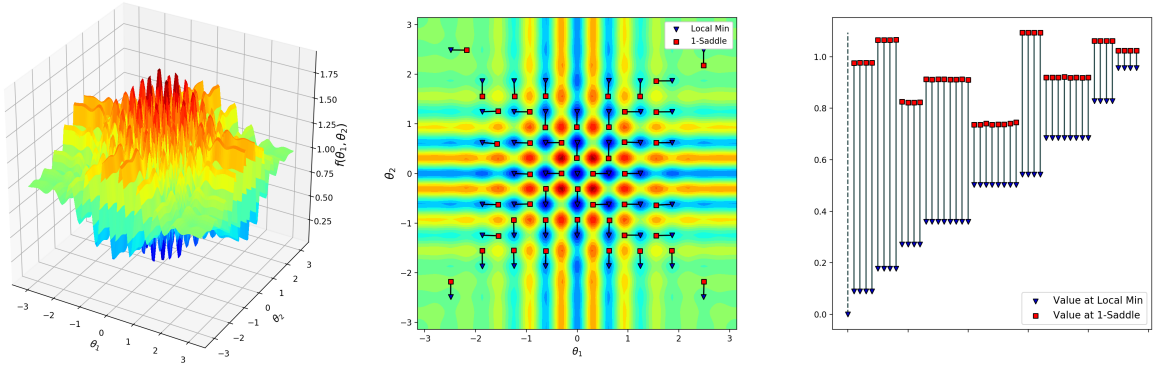
$$f(\theta_1, \theta_2) = 1 - \frac{1}{2}\sum_{d=1}^{2} \cos(10\,\theta_d) \cdot e^{-\frac{\theta_d^2}{2}},$$

   Wavy function is symmetric with respect to the dihedral group D_4 of order 8. Thus, its minimum-saddle critical value pairs come in multiplets forming simple representations of the group D_4.



(a) Surface plot (3D). (b) Correspondence between minima and 1-saddles (c) Barcode of local minima.

Figure 3: Langermann function, its minima-saddle correspondence and barcodes computed by Algorithm 1.



(a) Surface plot (3D). (b) Color plot (2D) & correspondence between minima and 1-saddle points. (c) Barcode of local minima.

Figure 4: Wavy function, its minima-saddles correspondence & barcode computed by Algorithm 1

4. **HolderTable** test objective function $f : [-10, 10]^2 \to \mathbb{R}$ (Figure 5):

$$f(\theta_1, \theta_2) = -\left| e^{|1-\frac{\sqrt{\theta_1^2+\theta_2^2}}{\pi}|} \sin(\theta_1)\cos(\theta_2) \right|$$

Convergence of the Algorithm

In this subsection we prove and test empirically the convergence of our algorithm when the number of points N used to construct barcodes tends to infinity. To compare two barcodes we adopt **Bottleneck distance** [20, 9], also known as Wasserstein-∞ distance $\mathbb{W}_\infty$, on the corresponding persistence diagrams (2-dimensional point clouds of birth-death pairs):

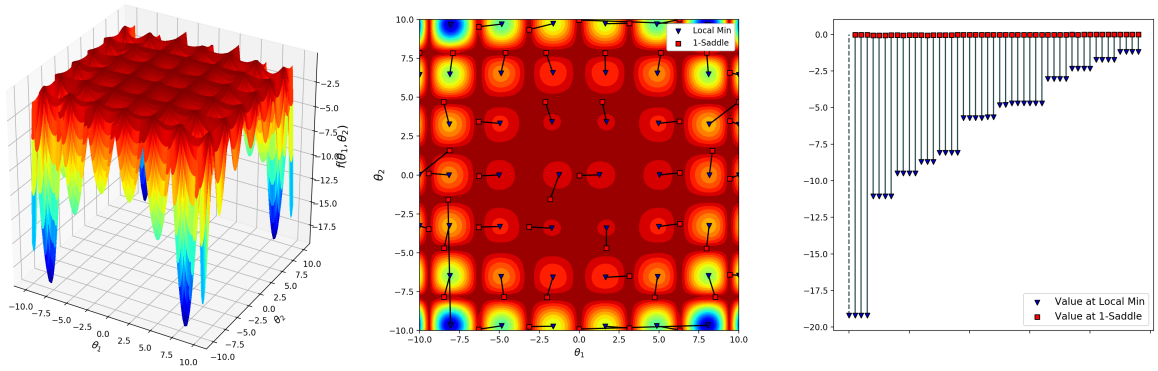
$$\mathbb{W}_\infty(\mathcal{D}, \mathcal{D}') := \inf_{\pi \in \Gamma(\mathcal{D}, \mathcal{D}')} \sup_{a \in \mathcal{D} \cup \Delta} |a - \pi(a)|. \tag{2}$$

Here Δ denotes the "diagonal" (pairs with birth equal to death) and $\Gamma(\mathcal{D}, \mathcal{D}')$ denotes the set of partial matchings between $\mathcal{D}$ and $\mathcal{D}'$ defined as bijections between $\mathcal{D} \cup \Delta$ and $\mathcal{D}' \cup \Delta$.

For a function $f : \Theta \to \mathbb{R}$ let $\mathcal{D}_f^*$ denote its barcode. Our algorithm uses finite number of N randomly sampled points $\Theta_N = \{\theta_1, \ldots, \theta_N\}$ to compute approximation $\hat{\mathcal{D}}_f(\Theta_N)$ of true barcode $\mathcal{D}_f^*$. Let C denote the Lipschitz constant of f

Proposition 5. *Let for any $\theta \in \Theta$ there exist $\theta_i \in \Theta_N$ such that $|\theta - \theta_i| < \varepsilon$. Then*

$$\mathbb{W}_\infty\left(\hat{\mathcal{D}}_f(\Theta_N), \mathcal{D}_f^*\right) < C\varepsilon$$



(a) Surface plot (3D). (b) Color plot (2D) & correspondence between minima and 1-saddle points. (c) Barcode of local minima.

Figure 5: HolderTable function, its minima-saddles correspondence and barcode computed by Algorithm 1

Proof. This follows for example from ([16], Lemma 1). □

Note that for typical sample Θ_N the maximal distance from $\theta \in \Theta$ to Θ_N is $\sim N^{-\frac{1}{D}}$. It follows that $\hat{\mathcal{D}}_f(\Theta_N)$ converges to $\mathcal{D}_f^*$ as $N \to \infty$ for any typical sequence of samples $\{\Theta_N\}$. And in particular

$$\mathbb{W}_\infty\big(\hat{\mathcal{D}}_f(\Theta_N), \hat{\mathcal{D}}_f(\Theta'_N)\big) \to 0 \tag{3}$$

as $N \to \infty$, for any typical pair of sequences $\{\Theta_N\}, \{\Theta'_N\}$.

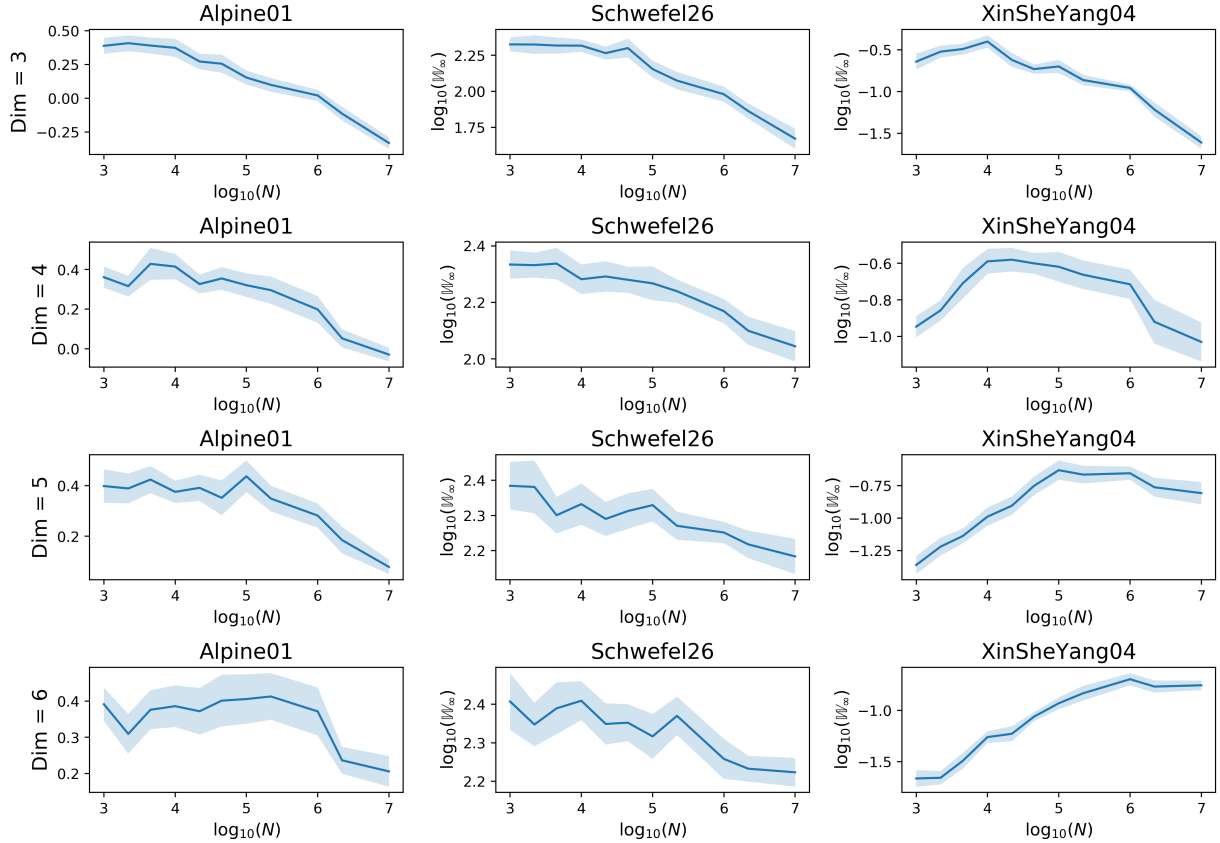


Figure 6: Decrease of logarithm of Bottleneck distance between pairs of persistent diagrams on samples of size N for Alpine01, Schwefel26, XinSheYang04 benchmark functions in dimensions $D \in \{3, 4, 5, 6\}$.

We have checked the condition (3) on several test function from Global Optimization benchmark. Each of the test function comes in a series of functions of arbitrary dimension. Even for small dimensions they are extremely complex (contain exponential number of local minima and saddle points). For each function f and dimension $D = 3, 4, 5, 6$, we consider various sample sizes $\log_{10} N \in [3, 3.5, 4, \dots, 7]$. For every triplet (f, D, N), we randomly sample $R = 20$ pairs $\big((\Theta_N)_r, (\Theta'_N)_r\big)$ of point clouds of size N. Next for each of $2R$ point clouds we compute the barcode by our algorithm and in each pair measure the bottleneck distance between the barcodes. Finally, we take the mean value

$$\mathbb{E}_{\Theta_N, \Theta'_N} \mathbb{W}_\infty\big(\hat{\mathcal{D}}_f(\Theta_N), \hat{\mathcal{D}}_f(\Theta'_N)\big) \approx \frac{1}{R} \sum_{r=1}^{R} \mathbb{W}_\infty\big(\hat{\mathcal{D}}_f((\Theta_N)_r), \hat{\mathcal{D}}_f((\Theta'_N)_r)\big). \tag{4}$$

The considered functions are:

1. **Alpine01** function $f : [-10, 10]^D \to \mathbb{R}$ defined by

$$f(\theta_1, \dots, \theta_D) = \sum_{d=1}^{D} |\theta_d \sin \theta_{x_d} + 0.1\theta_d|.$$

2. **Schwefel26** function $f : [-500, 500]^D \to \mathbb{R}$ defined by:

$$f(\theta_1, \dots, \theta_D) = 418.9829 \cdot D - \sum_{d=1}^{D} \theta_d \sin(\sqrt{|\theta_d|}).$$

3. **XinSheYang04** function $f : [-10, 10]^D \to \mathbb{R}$ defined by:

$$f(\theta_1, \dots, \theta_D) = \left[\sum_{d=1}^{D} \sin^2(\theta_d) - e^{-\sum_{d=1}^{D} \theta_d^2}\right] e^{-\sum_{d=1}^{D} \sin^2 \sqrt{|\theta_d|}}$$

For every pair (f, D) we sum up the dependence on cloud size N in a form of a plot in the decimal logarithmic scale. We observe empirically that for big enough N

$$\log(\mathbb{W}_\infty\big(\hat{\mathcal{D}}_f(\Theta_N), \hat{\mathcal{D}}_f(\Theta'_N)\big)) \sim -\frac{1}{D} \log(N)$$

as it is expected based on Proposition 5.

The results are summarized in Figure 6.

In Table 1 the running times for the Algorithm 1, including HNSW NN search, on random samples of 10^6 points, are compared with the running times for platform GUDHI on 10^6 points grid, for Alpine01 and Schwefel26 benchmark functions in dimensions $D \in \{3, 4, 5, 6, 7\}$.

	Alpine01	Schwefel26	Alpine01	Schwefel26	Alpine01	Schwefel26	Alpine01	Schwefel26	Alpine01	Schwefel26
	3D	3D	4D	4D	5D	5D	6D	6D	7D	7D
Gudhi	9	10	28	29	72	79	160	181	767	863
Alg 1	12	12	14	14	16	16	18	18	22	21

Table 1: Running time in seconds on 10^6 points, on Intel(R) Xeon(R) CPU E5-2698 v4 @2.20GHz

5. TOPOLOGY OF NEURAL NETWORKS' LOSS FUNCTIONS

Clearly some minima can harm more than others gradient-based algorithms. For example two minima on Figure 7 locally look the same but pose different obstacles for gradient-based optimization. A gradient-based algorithm trajectory can have difficulty to escape from a vicinity of local minimum. To reach a point with lower loss, a path, coming from a vicinity of a given local minimum, has to climb up to points with higher loss. The minimal value of this higher loss penalty is precisely the loss at 1-saddle associated with the minimum, from definition of barcode. The bigger the minimum's segment in the barcode the more difficulty a gradient-based learning trajectory has escaping vicinity of the minimum in order to reach lower loss points.

In this section we compute and analyse barcodes of small fully connected neural networks with up to three hidden layers.

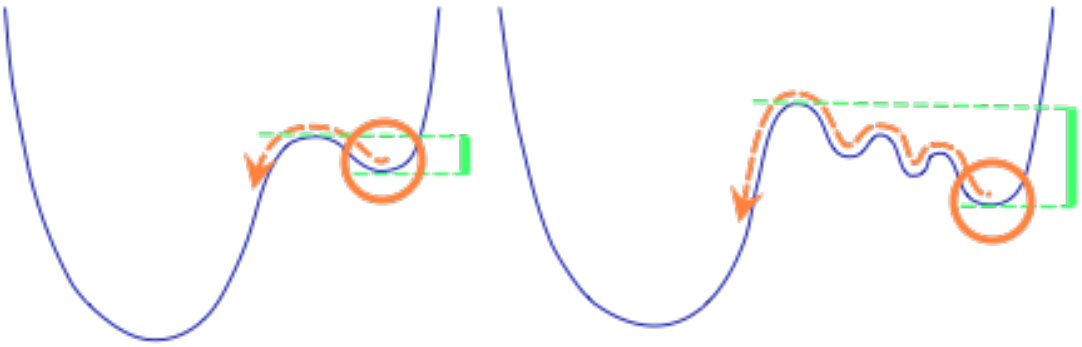


Figure 7: Barcodes quantify topological obstacles posed by local minima for gradient-based optimization. Here the two local minima locally look the same but clearly pose different obstacles for gradient-based learning. These obstacles are quantified by the lengths of green segments, which are associated with the local minima in the barcode.

For several architectures of the neural networks many results on the loss surface are known (see e.g. [21, 22] and references therein). Different geometrical and topological properties of loss surfaces were studied in [23, 24, 25, 26].

We have analyzed neural networks that are small. However our method permits full exploration of the loss surface as opposed to stochastical exploration of higher-dimensional loss surfaces. Let us emphasize that even from practical point of view it is important to understand first the behavior of barcodes in simplest examples where all hyper-parameters optimization schemes can be easily turned off.

For every analysed neural network the loss function is the mean squared error for predicting (randomly selected) function $g : [-\pi, \pi] \to \mathbb{R}$ given by

$$g(x) = 0.31 \cdot \sin(-x) - 0.72 \cdot \sin(-2x) - 0.21 \cdot \cos(x) + 0.89 \cdot \cos(2x)$$

plus l_2-regularization. The error is computed for prediction on uniformly distributed inputs $x \in \{-\pi + \frac{2\pi}{100}k \mid k = 0, 1, \ldots, 100\}$.

The neural networks considered were fully connected one-hidden layer with 2 and 3 neurons, two-hidden layers with 2x2, 3x2 and 3x3 neurons, and three hidden layers with 2x2x2 and 3x2x2 neurons with *ReLU* activation. We have calculated the barcodes of the loss functions on the hyper-cubical sets Θ which were chosen based on the typical range of parameters of minima. The results are as shown in Figure 8.

We summarize our findings into two main observations:

1. The barcodes are located in tiny lower part of the range of values; typically the maximum value of the function was around 200 and higher, and the saddles paired with minima lie well below 1;
2. With the increase of the neural network depth and width the barcodes descend lower.

For example the upper bounds of barcodes of one-layer (2) net are in range $[0.55, 0.65]$, two-layer (2×2) net in range $[0.35, 0.45]$, and three-layer $(2 \times 2 \times 2)$ net in range $[0.1, 0.3]$.

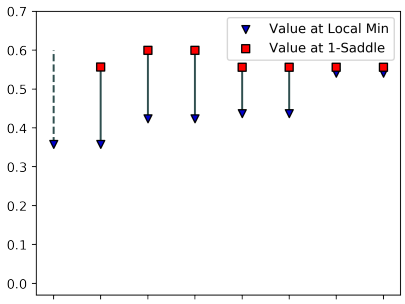
Implications for learning.

All minima in the observed cases are located in low part of loss function's range and they descend lower as the depth and the width of neural network increases. The gradient flow trajectories always have minima as terminal points for all but a subset of measure zero starting points. It follows that essentially any terminal point of gradient flow gives rather good solution to the regression problem. The precision of solution increases with increase of the neural network depth and width.

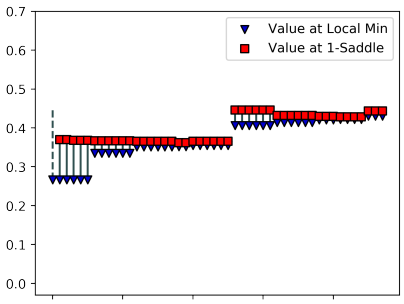
During learning, the gradient descent trajectory cannot get stuck at high local minima, since essentially all minima are located in a tiny low part of the function's range. There could perhaps exist some noise minima slightly higher but they are easily escaped during learning since their barcode is low, which implies that there always exists an escape path with low penalty.

If the length of barcode's segment is small for any local minimum then there always exists a small loss path to a lower minimum, this implies that gradient descent based optimization methods can in principle reach the lowest minimum during learning.

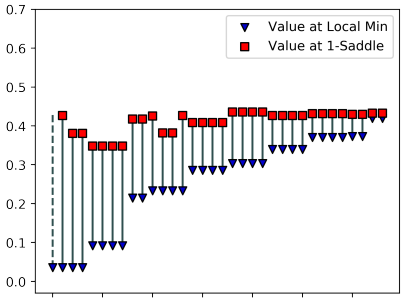
If the barcodes of all local minima are low this means also that any two such minima can be connected by a small loss path. This means that for any two local minima there exists continuous low loss transformation



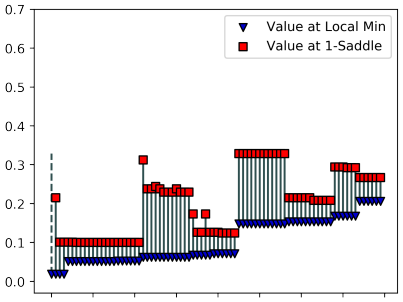
(a) Barcode for (2) net



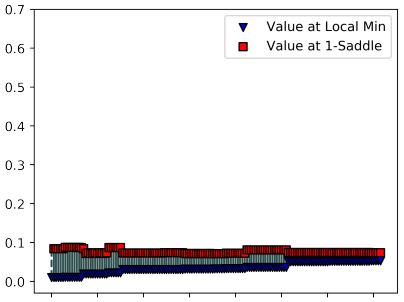
(b) Barcode for (3) net



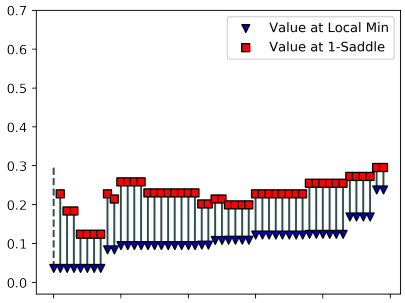
(c) Barcode for (2×2) net



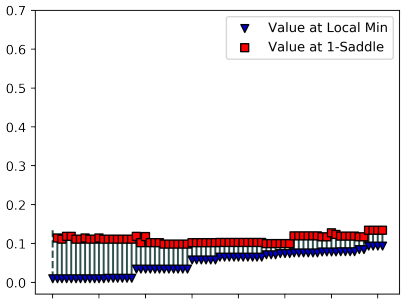
(d) Barcode for (3×2) net



(e) Barcode for (3×3) net



(f) Barcode for $(2 \times 2 \times 2)$ net



(g) Barcode for $(3 \times 2 \times 2)$ net

Figure 8: Barcodes of different neural network loss surfaces.

between predictions of one minimum to predictions of another minimum, implying that their predictions are somehow equivalent.

Strategy for computing barcodes for deep neural networks.

The exploration of loss surface using topological data analysis methods can be extended to deep neural networks. The strategy for deep neural networks is based on the proposition 2 above. The idea is to act by gradient descent on a path starting from the given minimum and going to a point with lower minimum. The deformation of the path under action of gradient flow is given by action of the component of the gradient which is orthogonal to the tangent direction of the path. The gradient component which is parallel to the path's tangent direction is absorbed into reparametrization of the path. Then the formula (1) gives an estimate for the critical value of "death" 1-saddle corresponding to this minimum.

Implications for generalization.

We have conducted some experiments that show that there exists a correlation between the height of barcodes of groups of low minima of same loss value and their generalization errors. Thus at least under certain mild conditions the lower the barcode for minima with same loss value the better their generalization properties.

6. CONCLUSION

In this work we have introduced a methodology for analysing the graphs of functions, in particular, loss surfaces of neural networks. The methodology is based on computing topological invariants called barcodes.

To compute barcodes we used a graph-based construction which approximates the function. Then we apply the algorithm we developed to compute the function's barcodes of local minima. Our experimental results of computing barcodes for small neural networks lead to two principal observations.

First all barcodes sit in a tiny lower part of the total function's range. Secondly the barcodes descend lower as the depth and width of neural network increases. From the practical point of view, this means that gradient descent optimization cannot get stuck in high local minima, and it is also not difficult to get from one local minimum to another (with smaller value) during learning.

The method that we developed has several further research directions. Although we tested the method on small neural networks, it is possible to apply it to large-scale modern neural networks such as convolutional networks (i.e. ResNet, VGG, AlexNet, U-Net, see [27]) for image-processing based tasks. However, in this case the graph-based approximation that we use requires wise choice of representative graph vertices as, dense filling of area by points is computationally intractable. There are clearly also connections, deserving further investigation, between the barcodes of local minima and optimal learning rates or the rates of convergence during learning.

This work was partially supported by RFBR grant 21-51-12005 NNIO_a

APPENDIX

A. GRADIENT MORSE COMPLEX

The gradient Morse complex (C_*, ∂_*), is defined as follows. For generic f the critical points p_α, $df\mid_{T_{p_\alpha}} = 0$, are isolated. Near each critical point p_α f can be written as $f = \sum_{l=1}^{j} -(x^l)^2 + \sum_{l=j}^{n}(x^l)^2$ in some local coordinates. The index of the critical point is defined as the dimension of the set of downward pointing directions at that point, or of the negative subspace of the Hessian:

$$\mathrm{index}(p_\alpha) = j$$

Then define

$$C_j = \oplus_{\mathrm{index}(p_\alpha)=j} \left[p_\alpha, \mathrm{or}(T^-_{p_\alpha})\right]$$

where or is an orientation on a negative subspace $T_{p_\alpha} = T^-_{p_\alpha} \oplus T^+_{p_\alpha}$ of the Hessian $\partial^2 f$.

Let

$$\mathcal{M}(p_\alpha, p_\beta) = \left\{\gamma : \mathbb{R} \to M^n \mid \dot\gamma = -(\mathrm{grad}_g f)(\gamma(t)), \lim_{t\to-\infty} = p_\alpha, \lim_{t\to+\infty} = p_\beta\right\}/\mathbb{R}$$

be the set of gradient trajectories connecting critical points p_α and p_β, where the natural action of $\mathbb{R}$ is by the shift $\gamma(t) \mapsto \gamma(t+\tau)$.

If $\mathrm{index}(p_\beta) = \mathrm{index}(p_\alpha) - 1$ then generically the set $\mathcal{M}(p_\alpha, p_\beta)$ is finite. Let

$$\#\mathcal{M}([p_\alpha, \mathrm{or}], [p_\beta, \mathrm{or}])$$

denote in this case the number of the trajectories, counted with signs taking into account a choice of orientation, between critical points p_α and p_β.

The linear operator ∂_j is defined by

$$\partial_j [p_\alpha, \mathrm{or}] = \sum_{\mathrm{index}(p_\beta)=j-1} [p_\beta, \mathrm{or}] \, \#\mathcal{M}(p_\alpha, p_\beta)$$

The description of the critical points on manifold Θ with nonempty boundary $\partial\Theta$ is modified slightly in the following way. A connected component of sublevel set is born also at a local minimum of restriction of f to the boundary $f\mid_{\partial\Theta}$, if $\mathrm{grad} f$ is pointed inside manifold Θ. The merging of two connected components can also happen at 1-saddle of $f\mid_{\partial\Theta}$, if $\mathrm{grad} f$ is pointed inside Θ. When we speak about minima and 1-saddles, this also means such critical points of $f\mid_{\partial\Theta}$. Similarly the set of generators of index j chains in Morse complex includes index j critical points of $f\mid_{\partial\Theta}$ with $\mathrm{grad} f$ pointed inside Θ. The differential is also modified similarly to take into account trajectories involving such critical points.

B. PROOF OF THE THEOREM 3

Theorem. ([7], Section 2) *Any $\mathbb{R}-$filtered chain complex C_* over field $\mathtt{k}$ can be brought by a linear transformation preserving the $\mathbb{R}-$filtration to "canonical form", a canonically defined direct sum of indecomposable $\mathbb{R}-$filtered complexes of two types:*

- *1-dimensional $\mathbb{R}-$filtered complex with trivial differential,* $\partial \tilde{e}_i^{(j)} = 0$, $\left\langle \tilde{e}_i^{(j)} \right\rangle = F_{\leqslant r}$, $r \in \mathbb{R}$,
- *2-dimensional $\mathbb{R}-$filtered complex with trivial homology* $\partial \tilde{e}_{i_2}^{(j+1)} = \tilde{e}_{i_1}^{(j)}$, $\left\langle \tilde{e}_{i_1}^{(j)} \right\rangle = F_{\leqslant s_1}$, $\left\langle \tilde{e}_{i_1}^{(j)}, \tilde{e}_{i_2}^{(j+1)} \right\rangle = F_{\leqslant s_2}$, s_1, $s_2 \in \mathbb{R}$.

The resulting canonical form is unique.

Proof. ([7], Section 2)

Let $\{e_i^{(n)}\}$ be a basis in the vector spaces C_n compatible with the filtration, so that each subspace $F_r C_n$ is the span $\left\langle e_1^{(n)}, \dots, e_{i_r}^{(n)} \right\rangle$. Notice that the filtration defines the natural order on the set of basis elements.

Let $\partial e_l^{(n)}$ have the required form for $n = j$ and $l \leqslant i$, or $n < j$ and all l. I.e. either $\partial e_l^{(n)} = 0$ or $\partial e_l^{(n)} = e_{m(l)}^{(n-1)}$, where $m(l) \neq m(l')$ for $l \neq l'$.

Let

$$\partial e_{i+1}^{(j)} = \sum_k e_k^{(j-1)} \alpha_k.$$

Let's move all the terms with $e_k^{(j-1)} = \partial e_q^j$, $q \leqslant i$, from the right to the left side. We get

$$\partial (e_{i+1}^{(j)} - \sum_{q \leqslant i} e_q^{(j)} \alpha_{k(q)}) = \sum_k e_k^{(j-1)} \beta_k$$

If $\beta_k = 0$ for all k, then define

$$\tilde{e}_{i+1}^{(j)} = e_{i+1}^{(j)} - \sum_{q \leqslant i} e_q^{(j)} \alpha_{k(q)},$$

so that

$$\partial \tilde{e}_{i+1}^{(j)} = 0,$$

and $\partial e_l^{(n)}$ has the required form for $l \leqslant i+1$ and $n = j$, and for $n < j$ and all l.

Otherwise let k_0 be the maximal k with $\beta_k \neq 0$. Then

$$\partial (e_{i+1}^{(j)} - \sum_{q \leqslant i} e_q^{(j)} \alpha_{k(q)}) = e_{k_0}^{(j-1)} \beta_{k_0} + \sum_{k < k_0} e_k^{(j-1)} \beta_k,$$

$\beta_{k_0} \neq 0$. Define

$$\tilde{e}_{i+1}^{(j)} = \left(e_{i+1}^{(j)} - \sum_{q \leqslant i} e_q^{(j)} \alpha_{k(q)} \right) / \beta_{k_0},$$

$$\tilde{e}_{k_0}^{(j-1)} = e_{k_0}^{(j-1)} + \sum_{k < k_0} e_k^{(j-1)} \beta_k / \beta_{k_0}.$$

Then

$$\partial \tilde{e}_{i+1}^{(j)} = \tilde{e}_{k_0}^{(j-1)}$$

and for $n = j$ and $l \leqslant i+1$, or $n < j$ and all l, $\partial e_l^{(n)}$ has the required form. If the complex has been reduced to "canonical form" on subcomplex $\oplus_{n \leqslant j} C_n$, then reduce similarly $\partial e_1^{(j+1)}$ and so on.

Uniqueness of the canonical form follows essentially from the uniqueness at each previous step. Let $\left\{ a_i^{(j)} \right\}$, $\left\{ b_i^{(j)} = \sum_{k \leqslant i} a_k^{(j)} \alpha_k \right\}$ be two bases of C_* for two different canonical forms. Assume that for all indexes $p < j$ and all n, and $p = j$ and $n \leqslant i$ the canonical forms agree. Let $\partial a_{i+1}^{(j)} = a_m^{(j-1)}$ and $\partial b_{i+1}^{(j)} = b_l^{(j-1)}$ with $m > l$, $a_m^{(j-1)}$ is not in the filtration subspace corresponding to $b_l^{(j-1)}$.

It follows that

$$\partial \left(\sum_{k \leqslant i+1} a_k^{(j)} \alpha_k \right) = \sum_{n \leqslant l} a_n^{(j-1)} \beta_n,$$

where $\alpha_{i+1} \neq 0$, $\beta_l \neq 0$. Therefore

$$\partial a_{i+1}^{(j)} = \sum_{n \leqslant l} a_n^{(j-1)} \beta_n / \alpha_{i+1} - \sum_{k \leqslant i} \partial a_k^{(j)} \alpha_k / \alpha_{i+1}.$$

On the other hand $\partial a_{i+1}^{(j)} = a_m^{(j-1)}$, with $m > l$, and $\partial a_k^{(j)}$ for $k \leqslant i$ are either zero or some basis elements $a_n^{(j-1)}$ different from $a_m^{(j-1)}$. This gives a contradiction.

Similarly if $\partial b_{i+1}^{(j)} = 0$, then

$$\partial a_{i+1}^{(j)} = -\sum_{k \leqslant i} \partial a_k^{(j)} \alpha_k / \alpha_{i+1}$$

which again gives a contradiction by the same arguments. This proves the uniqueness of the canonical form. □

Remark 2. *The barcode of $\mathbb{R}-$filtered chain complex consists of segments representing the indecomposable $\mathbb{R}-$filtered chain complexes, see Definition 2 in Section 2. There is the standard lexicographic order on a set of such segments. The direct sum from the theorem statement is the standard "direct sum over a set" vector space.*

REFERENCES

[1] H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein, Visualizing the loss landscape of neural nets, in: Advances in Neural Information Processing Systems, 2018, pp. 6389–6399.

[2] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, Advances in Neural Information Processing Systems 27 (2014) 2933–2941.

[3] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, Y. LeCun, The loss surfaces of multilayer networks, JMLR Workshop and Conference Proceedings 38 (2015).

[4] R. Bott, Lectures on Morse theory, old and new, Bulletin of the american mathematical society 7 (2) (1982) 331–358.

[5] S. Smale, Differentiable dynamical systems, Bulletin of the American mathematical Society 73 (6) (1967) 747–817.

[6] R. Thom, Sur une partition en cellules associée à une fonction sur une variété, Comptes Rendus de l'Academie des Sciences 228 (12) (1949) 973–975.

[7] S. Barannikov, Framed Morse complexes and its invariants., Advances in Soviet Mathematics 21 (1994) 93–116. doi:10.1090/advsov/021/03.

[8] D. Le Peutrec, F. Nier, C. Viterbo, Precise Arrhenius law for p-forms: The Witten Laplacian and Morse–Barannikov complex, Annales Henri Poincaré 14 (3) (2013) 567–610. doi:10.1007/s00023-012-0193-9.

[9] F. Le Roux, S. Seyfaddini, C. Viterbo, Barcodes and area-preserving homeomorphisms, Geometry & Topology 25 (6) (2021) 2713–2825.

[10] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of Machine Learning Research 13 (Feb) (2012) 281–305.

[11] P. B. M. K. Chung, P. T. Kim, Persistence diagrams of cortical surface data, Information Processing in Medical Imaging 5636 (2009) 386–397.

[12] T. Sousbie, C. Pichon, H. Kawahara, The persistent cosmic web and its filamentary structure – II. Illustrations, Monthly Notices of the Royal Astronomical Society 414 (1) (2011) 384–403. doi:10.1111/j.1365-2966.2011.18395.x.

[13] C. S. Pun, K. Xia, S. X. Lee, Persistent-homology-based machine learning and its applications – a survey, preprint arxiv: 1811.00252 (2018). arXiv:1811.00252.

[14] C. Dellago, P. G. Bolhuis, P. L. Geissler, Transition Path Sampling, John Wiley & Sons, Ltd, 2003, pp. 1–78. doi:10.1002/0471231509.ch1.

[15] A. R. Oganov, M. Valle, How to quantify energy landscapes of solids, The Journal of Chemical Physics 130 (10) (2009) 104504. doi:10.1063/1.3079326.

[16] F. Chazal, L. Guibas, S. Oudot, P. Skraba, Scalar field analysis over point cloud data, Discrete & Computational Geometry 46 (4) (2011) 743.

[17] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, Stability of persistence diagrams, Discrete & Computational Geometry 37 (1) (2007) 103–120.

[18] Y. A. Malkov, D. A. Yashunin, Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs, IEEE transactions on pattern analysis and machine intelligence (2018).

[19] M. Jamil, X.-S. Yang, A literature survey of benchmark functions for global optimization problems, International Journal of Mathematical Modelling and Numerical Optimisation 4 (2) (2013) 150–194.

[20] A. Efrat, A. Itai, M. J. Katz, Geometry helps in bottleneck matching and related problems, Algorithmica 31 (1) (2001) 1–28.

[21] K. Kawaguchi, Deep learning without poor local minima, in: Advances in neural information processing systems, 2016, pp. 586–594.

[22] M. Gori, A. Tesi, On the problem of local minima in backpropagation, IEEE Transactions on Pattern Analysis & Machine Intelligence 14 (1) (1992) 76–86.

[23] J. Cao, Q. Wu, Y. Yan, L. Wang, M. Tan, On the flatness of loss surface for two-layered relu networks, in: Asian Conference on Machine Learning, 2017, pp. 545–560.

[24] M. Yi, Q. Meng, W. Chen, Z.-m. Ma, T.-Y. Liu, Positively scale-invariant flatness of relu neural networks, arXiv:1903.02237 (2019).

[25] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, R. Zecchina, Entropy-sgd: Biasing gradient descent into wide valleys, in: International Conference on Learning Representations (ICLR), 2017.

[26] L. Dinh, R. Pascanu, S. Bengio, Y. Bengio, Sharp minima can generalize for deep nets, in: Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, PMLR, 2017, pp. 1019–1028.

[27] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, M. Hasan, B. C. Van Essen, A. A. Awwal, V. K. Asari, A state-of-the-art survey on deep learning theory and architectures, Electronics 8 (3) (2019) 292.

Keywords: loss surface, persistent homology, persistence barcodes, Morse theory, neural networks