Dazibao on picewise linear approximations

by Andrei Agrachev

Abstract

This text on piecewise linear continuous approximations of real functions doesn't pretend to be a regular mathematical paper, this is another genre. I call it dazibao.

The motivation for this dazibao is the extremely popular and successful way to “learn” functions by “training deep neural networks”.

Let us try to “learn” a real function of d variables from sample values of the function at the points of a fixed finite subset $Z \subset \mathbb{R}^d$. To this end, we fix a class $\mathcal{F}$ of available functions and try to approach $\min\limits_{\varphi \in \mathcal{F}} J(\varphi)$,

$$J(\varphi) = \sum_{\zeta \in Z} \rho(\varphi(\zeta) - y_\zeta),$$

where ρ is a penalty function (a convex function such that $\rho(-t) = \rho(t)$, $\rho(0) = 0$) and y_ζ is the expected value of the function which we are learning at the point ζ. A minimizer of J is the best we can learn with our data.

This is a quite common approximation problem. In classical mathematics, $\mathcal{F}$ is usually a finite-dimensional vector space of functions. It happens however that bulky nonlinear objects called “deep neural networks” are incomparably more efficient in a huge number of practical applications.

Deep neural networks is nowadays a colossal business but their simple basic incarnations are as follows. A function φ is obtained by a d-layers neural network if it has a form:

$$\varphi(x) = \langle a^0, F^1 \circ \cdots \circ F^m(x) \rangle + \alpha^0, \quad x \in \mathbb{R}^d, \tag{1}$$

where $F^i : \mathbb{R}^{k_{i-1}} \to \mathbb{R}^{k_i}$,

$$F^i(x) = \left(\sigma(\langle a_1^i, x\rangle + \alpha_1^i), \ldots, \sigma(\langle a_{k_i}^i, x\rangle + \alpha_{k_i}^i)\right), \quad x \in \mathbb{R}^{k_i - 1}, \tag{2}$$

$a^i_j \in \mathbb{R}^{k_i-1}$, $\alpha^j_i \in \mathbb{R}$, $i = 0, \dots, m$, and

$$\sigma(y) = \max\{\varepsilon y, y\}, \quad y \in \mathbb{R}, \tag{3}$$

where $\varepsilon \in [0, 1)$ is a fixed in advance constant. Here σ is the "activating function", the only nonlinear element of the construction.

Function (1)-(3) is continuous and "piecewise linear". A simple way to rigorously characterize continuous piecewise linear functions is to define them as continuous selections of linear or affine ones. Let $e_1, \cdots, e_n$ be affine functions on $\mathbb{R}^d$, $e_i(x) = s^0_i + \sum\limits_{j=1}^{d} s^j_i x_j$, $s^j_i, x_j \in \mathbb{R}$. We say that a continuous function φ is continuous selection of the family $e_1, \dots e_n$ if

$$\varphi(x) \in \{e_1(x), \dots, e_l(x)\}, \quad \forall\, x \in \mathbb{R}^d.$$

Minimal number of the affine functions in the family is a natural measure of complexity of the piecewise linear function φ.

Any continuous piecewise linear function should admit a realization by a neural network of the form (1)-(3) if we do not have restrictions on the number or dimensions of the layers; these numbers somehow correlate with the complexity of φ. Fairly, such a realization of a piecewise linear function looks rather artificial and very hard to study. The next example should make more clear what I mean.

Let us substitute the activating function $\sigma(y) = \max\{\varepsilon y, y\}, y \in \mathbb{R}$, in (2) by the function $\hat{\sigma}(y) = y^2$. Such a modification of the neural network (1)-(3) produces a polynomial function $\hat{\varphi}$. Moreover, it is easy to see that any polynomial can be realized in this way. Now try to recognize properties of the polynomial from such a neural network type realization!

Actually, there is a much more simple and transparent presentation of continuous piecewise linear functions found in [1]. Recall that an abstract simplicial complex with n vertices is a collection of subsets $S_i \subset \{1, \dots, n\}$, $i = 1, \dots, k$ such that $S_i \not\subseteq S_j$ if $i \neq j$.

Theorem 1. *Let φ be a continuous selection of the affine functions $e_1, \dots, e_n$. Then there exists an abstract simplicial complex $S_1, \dots, S_k$ with n vertices such that*

$$\varphi(x) = \min\Big\{\max_{j \in S_1} e_j(x), \dots, \max_{j \in S_k} e_j(x)\Big\}. \tag{4}$$

Moreover, if $e_1, \dots, e_n$ are affinely independent then this abstract simplicial complex is unique.

It is convenient to think about the affine function $e(x) = s^0 + \sum_{j=1}^{d} s^j x_j$ on $\mathbb{R}^d$ as a linear function $\sum_{j=0}^{d} s^j x_j$ on $\mathbb{R}^{1+d}$ restricted to the affine hyperplane defined by the equation $x_0 = 1$. The affine independence of e_j is just linear independence of the vectors $(s_j^0, s_j^1, \ldots, s_j^d)$, $1 \le j \le n$.

Let Σ_n be the group of permutations of $\{1, \ldots, n\}$. For any $\nu \in \Sigma_n$, we set

$$K_\nu = \{x \in \mathbb{R}^d : e_{\nu(1)}(x) \le \cdots \le e_{\nu(n)}(x)\};$$

then K_ν is a convex polytope (not necessary bounded and maybe empty). It is easy to see that φ is linear on K_ν, there exists $\phi(\nu) \in \{1, \ldots, n\}$ such that $\varphi\big|_{K_\nu} = e_{\nu(\phi(\nu))}\big|_{K_\nu}$. Indeed,

$$intK_\nu = \{x \in \mathbb{R}^d : e_{\nu(1)}(x) < \cdots < e_{\nu(n)}(x)\}.$$

If $intK_\nu \ne \emptyset$, then we do not have a chance to switch from one e_i to another one inside K_ν. If K_ν is not full-dimensional then we restrict everything to the affine hull of K_ν where it is full-dimensional. The restriction may lead to the loss of uniqueness of $\phi(\nu)$ because the restrictions of different e_i maybe equal.

Let $\bar{e} = (e_1, \ldots, e_n)$, $\Sigma_{\bar{e}} = \{\nu \in \Sigma_n : intK_\nu \ne \emptyset\}$; then $\mathbb{R}^d = \overline{\bigcup_{\nu \in \Sigma_{\bar{e}}} intK_\nu}$. We see that the continuous selection φ is uniquely determined by its symbol $\phi : \Sigma_{\bar{e}} \to \{1, \ldots, n\}$.

Theorem 1 follows from the following key lemma whose proof will be explained later.

Key lemma. *Let* $\nu \in \Sigma_{\bar{e}}$, $S_\nu = \{\nu(1), \ldots, \nu(\phi(\nu))\}$, *then*

$$\varphi(x) \ge \max_{j \in S_\nu} e_j(x), \quad \forall\, x \in \mathbb{R}^d.$$

Key lemma implies that $\varphi(x) = \min_{\nu \in \Sigma_{\bar{e}}} \max_{j \in S_\nu} e_j(x)$. Moreover, the min max does not change if we remove all sets of the collection S_ν, $\nu \in \Sigma_{\bar{e}}$, which have proper subsets in the same collection.

Remark. Inverting the order in Key lemma we obtain the inequality

$$\varphi(x) \le \max\left\{e_{\nu(\phi(\nu))}, \ldots, e_{\nu(n))}(x)\right\}, \quad \forall\, x \in \mathbb{R}^d,$$

and the max min presentation:

$$\varphi(x) = \max_{\nu \in \Sigma_{\bar{e}}} \min_{j \in S^{\nu}} e_j, \qquad \text{where } S^{\nu} = \{\nu(\phi(\nu)), \dots, \nu(n)\}.$$

Theorem 1 and its max min version are suitable both for the topological study of piecewise linear functions and for the "learning". We first explain topology, a piecewise linear analogue of the Morse theory.

We say that the n-tuple of affine functions $\bar{e} = (e_1, \dots, e_n)$ is in general position if any $d+1$ functions among $e_1, \dots, e_n$ are affinely independent. Clearly, n-tuples in general position form a Zariski-open subset of the $(d+1)n$-dimensional space of all n-tuples $\bar{e}$. Continuous selections of the n-tuples in general position play here the role of Morse functions.

Given $x \in \mathbb{R}^d$, we define the set of active indices $I_x(\varphi) \subset \{1, \dots, n\}$ by the formula:

$$I_x(\varphi) = \{\nu(\phi(\nu)) : x \in K_{\nu},\ \nu \in \Sigma_{\bar{e}}\}.$$

If $\bar{e}$ is in general position, then $\# I_x(\varphi) \le d+1,\ \forall\, x \in \mathbb{R}^d$.

Let $B^d = \{(y_1, \dots, y_d) \in \mathbb{R}^d : |y_i| < 1,\ i = 1, \dots, d\}$, a d-dimensional box. We say that $x \in \mathbb{R}^d$ is a topologically regular point of function φ, if there exists a neighborhood $O_x \subset \mathbb{R}^d$ of the point x and a homeomorphism $\Psi : B_{\varepsilon} \to O_x$ such that $\varphi \circ \Psi(y) = \varphi(x) + y_1,\ \forall\, y \in B^d$. Otherwise, x is a topologically critical point of φ.

Let $\bar{e}$ be in general position and φ be a continuous selection of $\bar{e}$. It is not hard to show that $\# I_x = d+1$ for any topologically critical point of φ. In particular, critical points are isolated. The homotopy type of the Lebesgue sets $\{x \in \mathbb{R}^d : \varphi(x) \le t\}$ and $\{x \in \mathbb{R}^d : \varphi(x) \ge t\}$ change when t passes a critical level of φ and these changes are controlled by the min max and max min presentations of φ.

We need some notations to describe the change of homotopy type. We do it in Theorems 2 and 3 where we assume that $\bar{e}$ is in general position and φ has form (4). We set $C_x^i = I_x(\varphi) \setminus (I_x \cap S_i),\ i = 1, \dots, k$. Maximal sets of the collection $C_x^i,\ i = 1, \dots, k$, with respect to the inclusion form a simplicial complex $\mathcal{C}_x$.

Let $\hat{\mathcal{C}}_x$ be the cone over $\mathcal{C}_x$; then $\mathcal{C}_x \subset \hat{\mathcal{C}}_x$ and $H_i(\hat{\mathcal{C}}_x; \mathcal{C}_x) = \tilde{H}_{i-1}(\mathcal{C}_x),\ i = 0, 1, \dots$, where $H_i(\cdot; \cdot)$ is the i-dimensional homology of the pair and $\tilde{H}_i(\cdot)$ is homology of the augmented chain complex. If $\mathcal{C}_x$ is contractible, then $H_i(\hat{\mathcal{C}}_x; \mathcal{C}_x) = 0,\ \forall\, i \ge 0$. If $\mathcal{C}_x$ is an empty complex, then $\hat{\mathcal{C}}_x$ is a point, $H_0(\hat{\mathcal{C}}_x; \emptyset) = \mathbb{Z},\ H_i(\hat{\mathcal{C}}_x; \emptyset), \forall\, i > 0$.

We set $\operatorname{cr}(\varphi) = \{x \in \mathbb{R}^d : \# I_x(\varphi) = d+1\}$.

Theorem 2. *For any $x \in \mathrm{cr}(\varphi)$ there exists a neighborhood O_x of x such that the pair $\big(O_x; \{y \in O_x : \varphi(y) < \varphi(x)\}\big)$ is homotopy equivalent to the pair $\big(\hat{\mathcal{C}}_x; \mathcal{C}_x\big)$.*

This theorem as well as the next one are special cases of a more general result proved in [2].

Theorem 3. *Assume that $\{x \in \mathbb{R}^d : \varphi(x) \leq t_1\}$ is compact and $t_0 < t_1$.*

If $\varphi^{-1}(t) \cap \mathrm{cr}(\varphi) = \emptyset\ \forall\, t \in (t_0, t_1]$, then $\{x \in \mathbb{R}^d : \varphi(x) \leq t_0\}$ is a homotopy retract of $\{x \in \mathbb{R}^d : \varphi(x) \leq t_1\}$.

If there exists exactly one $t \in (t_0, t_1]$ such that $\varphi^{-1}(t) \cap cr(\varphi) \neq \emptyset$, then

$$H_i\big(\{x \in \mathbb{R}^d : \varphi(x) \leq t_1\}; \{x \in \mathbb{R}^d : \varphi(x) \leq t_0\}\big) = \bigoplus_{x \in \mathrm{cr}(\varphi) \cap \varphi^{-1}(t)} \tilde{H}_{i-1}(\mathcal{C}_x).$$

This is what concerns topology. Now we turn to the learning and we do not assume anymore that $\bar{e}$ ia in general position. It is wise to use the term “learning” instead of “optimization”; the goal of an iterated process is to reasonably improve the approximation but we do not expect to really minimize the cost J. Moreover, as you will see, the procedure does not depend on the shape of the cost.

The class of functions $\mathcal{F}$ is one of piecewise linear continuous functions of a prescribed complexity. In other words, we fix n and consider all continuous selections of n affine functions. We start with a sample n-tuple $\bar{e} = (e_1, \dots, e_n)$ and try to select a collection $S_i \subset \{1, \dots, n\},\ i = 1, \dots, k$, such that the values of function (4) at the points $\zeta \in Z$ reasonably well approximate y_ζ.

How to find $S_1, \dots, S_k$ if we know that there exists their choice with the perfect match $\varphi(\zeta) = y_\zeta,\ \zeta \in Z$? The answer is given by the Key lemma. Indeed, for any $\zeta \in Z$, we take $\nu_\zeta \in \Sigma_{\bar{e}}$ such that $\zeta \in K_{\nu_\zeta}$ and take $\phi(\nu_\zeta) \in \{1, \dots, n\}$ such that $y_\zeta = e_{\nu_\zeta}(\phi(\nu_\zeta))$. Then we consider the family of subsets

$$\{\nu_\zeta(1), \dots, \nu_\zeta(\phi(\nu_\zeta))\}, \quad \zeta \in Z.$$

The required collection $S_1, \dots, S_k$ is formed by the minimal elements of this family with respect to the inclusion.

Of course, perfect match is an extremely rare luck. Anyway, I suggest to follow similar procedure in the most general case: for any $\zeta \in Z$, take $\nu_\zeta \in \Sigma_{\bar{e}}$ such that $\zeta \in K_{\nu_\zeta}$ and take $\phi(\nu_\zeta) \in \{1, \dots, n\}$ such that

$$|y_\zeta - e_{\nu_\zeta}(\phi(\nu_\zeta))| = \min_{1 \leq i \leq n} |y_\zeta - e_i(\zeta)|.$$

The desired $S_1, \dots, S_k$ are all minimal elements of the family of subsets $\{\nu_\zeta(1), \dots, \nu_\zeta(\phi(\nu_\zeta))\}$, $\zeta \in Z$.

As soon as $\mathcal{S} = \{S_1, \dots, S_k\}$ is fixed we consider

$$J\left(\min_{S\in\mathcal{S}} \max_{j\in S} e_j(\cdot)\right) \tag{5}$$

as a function of $\bar{e} = (e_1, \dots, e_n)$. "Learning" procedure which I am going to describe provides us with a new better n-tuple of affine functions $\bar{e}' = (e_1', \dots, e_n')$. Then we select new collection of subsets $S_i' \subset \{1, \dots, n\}$, $i = 1, \dots, k'$, based on the n-tuple $\bar{e}'$, fix it, "learn" new $\bar{e}''$, e. t. c.

Now we focus on the study of (5) as a function of $\bar{e}$ with a fixed simplicial complex $\mathcal{S} = \{S_1, \dots, S_k\}$. We improve $\bar{e}$ step by step. First we choose some $i \in \{1, \dots, n\}$ and change only e_i keeping all $e_j\, j \neq i$, fixed; then we choose another i and repeat, and so on.

It is reasonable to choose i randomly. We can start with the uniform probability distribution on $\{1, \dots, n\}$ such that each i has probability $\frac{1}{n}$ and we adjust the distribution after each step in such a way that recently involved i have smaller probabilities. A natural adjustment rule is as follows. Assume that number j has probability p_j at a certain step, $j = 1, \dots, n$. We randomly select i in this step and send $p_i \mapsto 0,\ p_j \mapsto p_j + \frac{p_i}{n-1}$, $\forall\, j \neq i$, for the next step. We stop iterations when we see that the learning cease to be efficient.

This kind of "random coordinate descend" should be well-known in the optimization theory. More interesting is "learning" of e_i when all e_j, $j \neq i$, are fixed. We do it in one step. Let me first give the formula and then explain why I believe that it is the right one.

Some notations. We may assume without lack of generality that the affine hull of Z is the whole $\mathbb{R}^d$, otherwise we restrict all our study to this affine hull. Given $\zeta \in Z$, we set $\tilde{\zeta} = (1, \zeta) \in \mathbb{R}^{1+d}$ and $\hat{\zeta} = \frac{1}{|\tilde{\zeta}|}\tilde{\zeta} \in \mathbb{S}^d$.

Let $\mathcal{R}_i = \{S \in \mathcal{S} : i \in S\}$. We define constants v_ζ^i, w_ζ^i by the following formulas:

$$v_\zeta^i = \min_{S\in\mathcal{R}_i} \max_{j\in S\setminus\{i\}} e_j(\zeta), \quad w_\zeta^i = \min_{S\in\mathcal{S}\setminus\mathcal{R}_i} \max_{j\in S} e_j(\zeta),$$

and set

$$u_\zeta^i = \frac{1}{|\tilde{\zeta}|} \min\bigl\{w_\zeta^i, \max\{y_\zeta, v_\zeta^i\}\bigr\}.$$

Finally, we identify affine function $e : x \mapsto s_0 + \sum\limits_{j=1}^{n} s_j x_j$ with vector $e = (s_0, s_1, \dots s_n) \in \mathbb{R}^{1+d}$; then $e(x) = \langle e, (1, x)\rangle$.

If $\mathcal{S}$ and all e_j, $j \neq i$, are fixed, then right e_i is going to be:

$$e_i = \left(\sum_{\zeta \in Z} \hat{\zeta}\hat{\zeta}^* \right)^{-1} \sum_{\zeta \in Z} u^i_\zeta \hat{\zeta}. \tag{6}$$

Why expression (6) is a good choice? We fix e_j, $j \neq i$, and try to approach minimum of the function

$$f : e_i \to J \left(\min_{S \in \mathcal{S}} \max_{j \in S} e_j(\cdot) \right).$$

We have, $f(e_i) = \sum\limits_{\zeta \in Z} \psi_\zeta(e_i)$, where

$$\psi_\zeta(e_i) = \rho \left(\min_{S \in \mathcal{S}} \max_{j \in S} e_j(\zeta) - y_\zeta \right) = \rho \left(\min\bigl\{ w^i_\zeta, \max\{e_i(\zeta), v^i_\zeta\} \bigr\} - y_\zeta \right).$$

Let $t = e_i(\zeta) = \langle e_i, \tilde{\zeta} \rangle$. We see that ψ_ζ is constant on the affine hyperplanes $t = const$. Moreover, ψ_ζ attains minimum at

$$t = \min\bigl\{ w^i_\zeta, \max\{y_\zeta, v^i_\zeta\} \bigr\} = |\tilde{\zeta}| u^i_\zeta.$$

In other words, ψ_ζ attains minimum at any point of the affine hyperplane

$$E_\zeta = \left\{ e \in \mathbb{R}^{1+d} : \langle e, \zeta \rangle = |\tilde{\zeta}| u^i_\zeta \right\}.$$

Now, instead of a search for the minimizer of $f = \sum\limits_{\zeta \in z} \psi_\zeta$, we simply look for the minimizer of the function

$$e \mapsto \frac{1}{2} \sum_{\zeta \in \ Z} dist\bigl(e, E_\zeta\bigr)^2, \quad e \in \mathbb{R}^{1+d},$$

which does not depend on the choice of the penalty function ρ.

Let us find this minimizer. We have, $dist(e, E_\zeta)^2 = \min\limits_{e_\zeta \in E_\zeta} |e - e_\zeta|^2$. Hence we are looking for solution of the following conditional minimum problem: minimize the function $(e, \{e_\zeta : \zeta \in Z\}) \mapsto \frac{1}{2} \sum_{\zeta \in Z} |e - e_\zeta|^2$ under conditions $e_\zeta \in E_\zeta,\ \zeta \in Z$.

We apply the Lagrange multipliers rule. The Lagrange function is:

$$L : (e, \{e_\zeta, \lambda_\zeta : \zeta \in Z\}) \mapsto \frac{1}{2} \sum_{\zeta \in Z} \bigl(|e - e_\zeta|^2 - \lambda_\zeta (\langle e, \tilde{\zeta} \rangle - y_\zeta) \bigr).$$

Condition $dL = 0$ gives the following equalities:

$$\sum_{\zeta \in Z} (e - e_\zeta) = 0, \quad e - e_\zeta = \lambda_\zeta \tilde{\zeta}, \quad \langle e_\zeta, \tilde{\zeta} \rangle = |\tilde{\zeta}| u^i_\zeta. \tag{7}$$

We take inner product of the 2nd equality in (7) with $\tilde{\zeta}$, apply the 3d equality and obtain: $\lambda_\zeta = \frac{\langle e, \tilde{\zeta} \rangle}{|\tilde{\zeta}|^2} - \frac{u^i_\zeta}{|\tilde{\zeta}|}$. Then we plugin this expression for λ_ζ in the second equality in (7) and get:

$$e - e_\zeta = \langle e, \hat{\zeta} \rangle \hat{\zeta} - u^i_\zeta \hat{\zeta}.$$

We sum up the last identity for all $\zeta \in Z$, use the 1st equality in (7) and obtain:

$$\left(\sum_{\zeta \in Z} \hat{\zeta} \hat{\zeta}^* \right) e = \sum_{\zeta \in Z} u^i_\zeta \hat{\zeta}. \tag{8}$$

Recall that the affine hull of $\zeta \in Z$ is the whole $\mathbb{R}^d$, hence the matrix in the left-hand side of (8) in nondegenerate and we obtain (6). Note that this matrix depends only on Z; it is the same for all iterations of the algorithm and has to be inverted only once.

The learning algorithm is completely described. It is hard to expect that it will work well if you use it blindly, as it is. Most probably, it will work very bad but it can serve as a base for further tuning.

To conclude, I explain the proof of the Key lemma, as promised. Let $\nu \in \Sigma_{\bar{e}}$, we can connect an interior point of K_ν with any point in $\mathbb{R}^d$ by a segment $x(t) = x(0) + t(x(1) - x(0))$, $0 \leq t \leq 1$, where $x(0) \in intK_\nu$. Let $u_i(t) = e_i(x(t))$; then $u_1(t), \ldots, u_n(t)$ are n numbered moving points on the real line, where each point is moving with its own constant velocity.

It follows that any point $u_i(t)$ can meet another point $u_j(t)$ only once during the movement. Moreover, $u_i(0) < u_j(0)$, $\forall\, i \in S_\nu$, $j \neq S_\nu$; hence $u_i(t)$ can meet $u_j(t)$ only in such a way that $u_i(t) < u_j(t)$ before the meeting and $u_i(t) > u_j(t)$ after the meeting.

For any $t \in [0, 1]$, one of the points $u_i(t)$ equals $\varphi(x(t))$; we mark this point by the checkbox. The checkbox can be transferred from one point to another point only when they meet. We thus obtain that, for any $t \in [0, 1]$, either the number of the marked point belongs to S_ν or some points whose numbers belong to S_ν are greater than the marked one. In particular, $\varphi(x(t)) \leq \max\limits_{i \in S_\nu} u_i(t)$, $\forall\, t \in [0, 1]$.

References

[1] S. Bartels, L. Kunz, S. Scholtes, *Continuous selections of linear functions and nonsmooth critical point theory.* Nonlinear Analysis, Theory, Methods & Applications, 1995, v.24, 385–407

[2] A. Agrachev, S. Scholtes, D. Pallaschke, *On Morse theory for piecewise smooth functions.* J. Dynamical and Control Systems, 1997, v.3, 449-469