A Universal Approximation Result
for Difference of log-sum-exp Neural Networks

Giuseppe C. Calafiore, *Fellow, IEEE*, Stephane Gaubert, *Member, IEEE*
and Corrado Possieri, *Associate Member, IEEE*

arXiv:1905.08503v1 [cs.NE] 21 May 2019

Abstract—We show that a neural network whose output is obtained as the difference of the outputs of two feedforward networks with exponential activation function in the hidden layer and logarithmic activation function in the output node (LSE networks) is a smooth universal approximator of continuous functions over convex, compact sets. By using a logarithmic transform, this class of networks maps to a family of subtraction-free ratios of generalized posynomials, which we also show to be universal approximators of positive functions over log-convex, compact subsets of the positive orthant. The main advantage of Difference-LSE networks with respect to classical feedforward neural networks is that, after a standard training phase, they provide surrogate models for design that possess a specific difference-of-convex-functions form, which makes them optimizable via relatively efficient numerical methods. In particular, by adapting an existing difference-of-convex algorithm to these models, we obtain an algorithm for performing effective optimization-based design. We illustrate the proposed approach by applying it to data-driven design of a diet for a patient with type-2 diabetes.

Index Terms—Feedforward neural networks, Universal approximation, LSE networks, Surrogate models, Subtraction-free expressions, DC programming, Data-driven optimization.

I. INTRODUCTION

A. Motivation

A well-known and compelling property of feedforward neural network (FFNN) models is that they are capable of approximating any continuous function over a compact set. Indeed, classical results in, e.g., [1], [2], show that, given any non-constant, bounded and continuous function, there exists a FFNN with a single hidden layer that can approximate it over a compact set. This *universal approximation* capability, together with the development of efficient algorithms to tune the network weights, paved the way to the efficient application of artificial neural networks in several frameworks, such as circuit design [3], control and identification of nonlinear systems [4], optimization over graphs [5], and many others.

However, when the goal of a neural network model is to construct a *surrogate model* for describing, and then optimizing, a complex input-output relation, it is of crucial importance that the structure of the model be well tailored for the subsequent numerical optimization phase. This is not usually the case for generic FFNN. Indeed, if the input-output model does not satisfy certain properties (such as, for instance,

convexity [6]), then designing the input so that the output is minimized, possibly under additional design constraints, can be an extremely difficult task.

In [7], we showed that LSE_T-networks, that is, FFNN with exponential activation functions in the inner layer and logarithmic activation function in the output neuron, parametrized by a positive "temperature" parameter $T > 0$, provide a smooth convex model capable of approximating any convex function over a convex, compact set. Maps in the LSE_T class are precisely log-Laplace transforms of nonnegative measures with finite support, a remarkable class of maps enjoying smoothness and strict convexity properties. We showed in particular that if the data to be approximated satisfy some convexity assumptions, such network structures can be readily exploited to perform data-driven design by using convex optimization tools. Nevertheless, most real-life input-output maps are of nonconvex nature, hence while they might still be approximated via a convex model, such an approximation may not yield a desirable accuracy.

B. Contribution

The purpose of this paper is to propose a new type of neural network model, here named *Difference-LSE network* (DLSE_T), which is constructed by taking the difference of the outputs of two LSE_T networks. First, we prove that DLSE_T networks guarantee universal approximation capabilities (thus overcoming the limitations of plain convex LSE_T networks), see Theorem 2. By using a logarithmic transformation, DLSE_T networks map to a family of ratios of generalized posynomials functions, which we show to be *subtraction free* universal approximators of positive functions over compact subsets of the positive orthant. Subtraction free expressions are fundamental objects in algebraic complexity, studied in particular in [8]. It is a result of independent interest that subtraction free expressions provide universal approximators.

Moreover, we show that Difference-LSE network are of practical interest, as they have a structure which is amenable to effective optimization over the inputs by using "DC-programming" methods, as discussed in Section VI.

Training and subsequent optimization of DLSE_T networks has been implemented in a numerical Matlab toolbox[1] named `DLSE_Neural_Network` that we made publicly available. The theoretical results in the paper are illustrated by an example dealing with data-driven design of a diet for a patient with type 2 diabetes.

G. C. Calafiore and C. Possieri are with the Dipartimento di Elettronica e Telecomunicazioni, Politecnico di Torino, 10129 Turin, Italy (e-mails: [giuseppe.calafiore, corrado.possieri]@polito.it). G.C. Calafiore is also with IEIIT-CNR Torino, 10129 Turin, Italy. S. Gaubert is with INRIA and CMAP, Ecole polytechnique, UMR 7641 CNRS, France (e-mail: stephane.gaubert@inria.fr).

[1]See https://github.com/Corrado-possieri/DLSE_neural_networks.

C. Related work

Similar to our previous work [7] on which it builds, the present paper is inspired by ideas from tropical geometry and max-plus algebra. The class of functions in LSE that we study here plays a key role in Viro's patchworking methods [9], [10] for real curves. We note that the application of tropical geometry to neural networks is an emerging topic: at least two recent works have used tropical methods to provide combinatorial estimates, in terms of Newton polytopes, of the "classifying power" of neural networks with piecewise affine functions, see [11], [12]. Other related results concern the "zero-temperature" ($T = 0$) limit of the approximation problem that we consider, i.e., the representation of piecewise linear functions by elementary expression involving min, max, and affine terms [13], [14], and the approximation of functions by piecewise linear functions. In particular, Th. 4.3 of [15] shows that any continuous function can be approximately arbitrarily well on a compact domain by a difference of piecewise linear convex functions. A related approximation result, still concerning differences of structured piecewise linear convex functions, has appeared in [16]. In contrast, our results provide approximation by differences of a family of *smooth* structured convex functions. This smoothing is essential when deriving universal approximation results by subtraction free rational expresssions.

II. Notation and technical preliminaries

Let $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$, $\mathbb{R}_{\geqslant 0}$, and $\mathbb{R}_{>0}$ denote the set of natural, integer, real, nonnegative real, and positive real numbers, respectively. Given a function $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, we define its domain as $\operatorname{dom} \phi \doteq \{\mathbf{x} \in \mathbb{R}^n : \phi(\mathbf{x}) < +\infty\}$. If the function $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is differentiable at point $\mathbf{x}$, we denote by $\nabla\phi(\mathbf{x})$ its gradient at $\mathbf{x}$.

A. Log-Sum-Exp functions

Following [7], we define LSE (Log-Sum-Exp) as the class of functions $f : \mathbb{R}^n \to \mathbb{R}$ that can be written as

$$f(\mathbf{x}) = \log\left(\sum_{k=1}^{K} b_k \exp(\langle \boldsymbol{\alpha}^{(k)}, \mathbf{x} \rangle)\right), \qquad (1)$$

for some $K \in \mathbb{N}$, $b_k \in \mathbb{R}_{>0}$, $\boldsymbol{\alpha}^{(k)} = [\ \alpha_1^{(k)} \ \cdots \ \alpha_n^{(k)}\ ]^\top \in \mathbb{R}^n$, $k = 1, \dots, K$, where $\mathbf{x} = [\ x_1 \ \cdots \ x_n\ ]^\top$ is a vector of variables. Further, given $T \in \mathbb{R}_{>0}$, we define LSE_T as the class of functions $f_T : \mathbb{R}^n \to \mathbb{R}$ that can be written as

$$f_T(\mathbf{x}) = T \log\left(\sum_{k=1}^{K} b_k^{1/T} \exp(\langle \boldsymbol{\alpha}^{(k)}, \mathbf{x}/T \rangle)\right), \quad (2)$$

for some $K \in \mathbb{N}$, $b_k \in \mathbb{R}_{>0}$, and $\boldsymbol{\alpha}^{(k)} \in \mathbb{R}^n$, $k = 1, \dots, K$. By letting $\beta_k \doteq \log b_k$, $k = 1, \dots, K$, we have that functions in the family LSE_T can be equivalently parameterized as

$$f_T(\mathbf{x}) = T \log\left(\sum_{k=1}^{K} \exp(\langle \boldsymbol{\alpha}^{(k)}, \mathbf{x}/T \rangle + \beta_k/T)\right), \quad (3)$$

where the β_ks have no sign restrictions. It may sometimes be convenient to highlight the full parameterization of f_T, in which case we shall write $f_T^{(\overrightarrow{\boldsymbol{\alpha}}, \boldsymbol{\beta})}$, where $\overrightarrow{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}^{(1)}, \dots, \boldsymbol{\alpha}^{(K)})$, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)$. It can be readily observed that, for any $T > 0$, the following property holds:

$$f_T^{(\overrightarrow{\boldsymbol{\alpha}}, \boldsymbol{\beta})}(\mathbf{x}) = T f_1^{(\overrightarrow{\boldsymbol{\alpha}}, \boldsymbol{\beta}/T)}(\mathbf{x}/T). \qquad (4)$$

The maps in LSE_T are special instances of the log-Laplace transforms of nonnegative measures, studied in [17]. In particular, the maps in LSE_T are smooth, they are convex (this is an easy consequence of Cauchy-Schwarz inequality), and they are even strictly convex if the vectors $\overrightarrow{\boldsymbol{\alpha}}^{(k)}$ constitute an affine generating family of $\mathbb{R}^n$, see [7, Prop. 1]. Maps of this kind play a key role in tropical geometry, in the setting of Viro's patchworking method [9], dealing with the degeneration of real algebraic curves to a piecewise linear limit. We note in this respect that the family of functions $(f_T)_{T>0}$ given by (3) converges uniformly on $\mathbb{R}^n$, as $T \to 0^+$, to the function

$$f_0(\mathbf{x}) \doteq \max_{1 \leqslant k \leqslant K} \left(\langle \boldsymbol{\alpha}^{(k)}, \mathbf{x} \rangle + \beta_k\right).$$

Actually, the following inequality holds for all $T > 0$

$$f_0(\mathbf{x}) \leqslant f_T(\mathbf{x}) \leqslant T \log K + f_0(\mathbf{x}), \qquad (5)$$

see [7] for details and background.

B. Posynomials and GPOS_T functions

Given $c_k > 0$ and $\boldsymbol{\alpha}^{(k)} \in \mathbb{R}^n$, a *positive monomial* is a product of the form $c_k \mathbf{x}^{\boldsymbol{\alpha}^{(k)}} = c_k x_1^{\alpha_1^{(k)}} x_2^{\alpha_2^{(k)}} \cdots x_n^{\alpha_n^{(k)}}$. A *posynomial* is a finite sum of positive monomials,

$$\psi(\mathbf{x}) = \sum_{k=1}^{K} c_k \mathbf{x}^{\boldsymbol{\alpha}^{(k)}}. \qquad (6)$$

We denote by POS the class of functions $\psi : \mathbb{R}_{>0}^n \to \mathbb{R}_{>0}$ of the form (6). Posynomials are *log-log-convex* functions, meaning that the log of a posynomial ψ is convex in the log of its argument, see, e.g., Section II.B of [7]. We denote by GPOS_T the class of functions that can be expressed as

$$\psi_T(\mathbf{x}) = (\psi(\mathbf{x}^{1/T}))^T \qquad (7)$$

for some $T > 0$ and $\psi \in \mathrm{POS}$. These functions are log-log-convex, and they form a subset of the so-called generalized posynomial functions, see, e.g., Section II.B of [7]. It is observed in Proposition 3 of [7] that LSE_T and GPOS_T functions are related by a one-to-one correspondence. That is, for any $f(\mathbf{x}) \in \mathrm{LSE}_T$ and $\psi(\mathbf{z}) \in \mathrm{GPOS}_T$ it holds that

$$\exp\left(f\left(\log(\mathbf{z})\right)\right) \in \mathrm{GPOS}_T, \quad \log\left(\psi\left(\exp(\mathbf{x})\right)\right) \in \mathrm{LSE}_T.$$

III. A universal approximation theorem

A. Preliminary: approximation of convex functions

We start by recalling a key result of [7], stating that functions in LSE_T are universal smooth approximators of convex functions, see Theorem 2 in [7].

Theorem 1 (Universal approximators of convex functions, [7]). *Let ϕ be a real valued continuous convex function defined*

on a compact convex subset $\mathcal{K} \subset \mathbb{R}^n$. Then, for all $\varepsilon > 0$ there exist $T > 0$ and a function $f_T \in \mathrm{LSE}_T$ such that

$$|f_T(\mathbf{x}) - \phi(\mathbf{x})| \leqslant \varepsilon, \quad \text{for all } \mathbf{x} \in \mathcal{K}. \tag{8}$$

We now extend the above result by showing that it actually holds for the restricted class of LSE_T with rational parameters. This extension will allow us to apply our results to the approximation by subtraction free expressions.

Definition 1. A function $f_T \in \mathrm{LSE}_T$ has *rational parameters* if $T > 0$ is a rational number and $f_T = f_T^{(\overrightarrow{\boldsymbol{\alpha}}, \boldsymbol{\beta})}$ is of the form (3) where the vectors $\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(K)}$ have rational entries, and $\beta_1, \ldots, \beta_K$ are rational numbers. We shall also say that f_T is a ε-approximation of f on $\mathcal{K}$ when (8) holds.

The following corollary holds.

Corollary 1 (LSE approximation with rational parameters). *Under the hypotheses of Theorem 1, for all $\varepsilon > 0$ there exists a rational $T > 0$ and a function $f_T \in \mathrm{LSE}_T$ with rational parameters such that* (8) *holds. Moreover, T may be chosen of the form $1/p$ where p is a positive integer.*

Proof. First, inspecting the proof of Theorem 2 in [7], one obtains that ϕ can be approximated uniformly by a map $f_T \in \mathrm{LSE}_T$, for all $T > 0$ small enough, hence we can always assume that T is of the form $1/p$ for some positive integer p. It then remains to be proved that the approximation result still holds if also $\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(K)}$ and $\beta_1, \ldots, \beta_K$ are rational. To this end, let us study the effect of a perturbation of these parameters on the map f_T. Observe that the map $\varphi : \mathbb{R}^K \to \mathbb{R}$, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K) \mapsto T \log(\sum_{k=1}^K \exp(\xi_k/T))$ satisfies

$$|\varphi(\boldsymbol{\xi}) - \varphi(\boldsymbol{\xi}')| \leqslant \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_\infty \tag{9}$$

for all $\boldsymbol{\xi} \in \mathbb{R}^K$, where $\|\cdot\|_\infty$ is the sup-norm. This follows from the fact that φ is order preserving and commutes with the addition of a constant, see e.g. [18] and Section 2 of [19]. It follows from (9) that if f_T is as in (3), and if

$$g_T(\mathbf{x}) = T \log \left(\sum_{k=1}^K \exp \left(\langle \boldsymbol{\gamma}^{(k)}, \mathbf{x}/T \rangle + \delta_k/T \right) \right),$$

then, letting $R \doteq \max_{\mathbf{x} \in \mathcal{K}} \|\mathbf{x}\|$, $\overrightarrow{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(K)})$, and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)$, we get

$$|f_T(\mathbf{x}) - g_T(\mathbf{x})| \leqslant \kappa((\overrightarrow{\boldsymbol{\alpha}}, \boldsymbol{\beta}), (\overrightarrow{\boldsymbol{\gamma}}, \boldsymbol{\delta}))$$
$$\doteq \max_{1 \leqslant k \leqslant K} \|\boldsymbol{\alpha}^{(k)} - \boldsymbol{\gamma}^{(k)}\| R + \max_{1 \leqslant k \leqslant K} |\beta_k - \delta_k|,$$

for all $\mathbf{x} \in \mathcal{K}$. Hence, choosing $(\overrightarrow{\boldsymbol{\gamma}}, \boldsymbol{\delta})$ to be a rational approximation of $(\overrightarrow{\boldsymbol{\alpha}}, \boldsymbol{\beta})$ such that $\kappa((\overrightarrow{\boldsymbol{\alpha}}, \boldsymbol{\beta}), (\overrightarrow{\boldsymbol{\gamma}}, \boldsymbol{\delta})) \leqslant \varepsilon$, and supposing that f_T is a ε-approximation of ϕ on $\mathcal{K}$, we deduce that g_T is a 2ε-approximation of ϕ on $\mathcal{K}$, from which the statement of the corollary follows. $\square$

B. Approximation of general continuous functions

This section contains our main result on universal approximation of continuous functions. To this end, we first define the class of functions that can be expressed as the difference of two functions in LSE_T.

Definition 2 (DLSE_T functions). We say that a function $\phi : \mathbb{R}^n \to \mathbb{R}$ belongs to the DLSE_T class, if $\phi = g_T - h_T$, for some $g_T, h_T \in \mathrm{LSE}_T$. Further, we say that ϕ has rational parameters, if g_T and h_T have rational parameters.

The following result shows that any continuous function can be approximated uniformly by a function in a DLSE_T class.

Theorem 2 (Universal approximation property of DLSE_T). *Let ϕ be a real-valued continuous function defined on a compact, convex subset $\mathcal{K} \subset \mathbb{R}^n$. Then, for any $\varepsilon > 0$ there exist a function $f_T \in \mathrm{DLSE}_T$ with rational parameters, for some $T = 1/p$ where p is a positive integer, such that $|f_T(\mathbf{x}) - \phi(\mathbf{x})| \leqslant \varepsilon$, $\forall \mathbf{x} \in \mathcal{K}$.*

Proof. A classical result of convex analysis states that any continuous function ϕ defined on a compact convex subset $\mathcal{K}$ of $\mathbb{R}^n$ can be written as the difference $g - h$ where g, h are continuous, convex functions defined on $\mathcal{K}$, see, e.g., Proposition 2.2 of [20]. Then, by Corollary 1, for all $\varepsilon > 0$, we can find a rational $T' > 0$ and a function $g_{T'} \in \mathrm{LSE}_{T'}$ with rational parameters such that $|g(\mathbf{x}) - g_{T'}(\mathbf{x})| \leqslant \varepsilon/2$ holds for all $\mathbf{x} \in \mathcal{K}$. Similarly, we can find a rational $T'' > 0$ and a function $h_{T''} \in \mathrm{LSE}_{T''}$ with rational parameters such that $|h(\mathbf{x}) - h_{T''}(\mathbf{x})| \leqslant \varepsilon/2$ holds for all $\mathbf{x} \in \mathcal{K}$. Hence, by taking any rational $T > 0$ such that T' and T'' are integer multiples of T, it follows from the nesting property in Lemma 1 of [7] that $g_{T'}$ and $h_{T''}$ both belong to LSE_T. Thus, there exist a rational $T > 0$ and $g_T, h_T \in \mathrm{LSE}_T$ such that, for all $\mathbf{x} \in \mathcal{K}$,

$$|g(\mathbf{x}) - g_T(\mathbf{x})| \leqslant \varepsilon/2 \qquad |h_T(\mathbf{x}) - h(\mathbf{x})| \leqslant \varepsilon/2.$$

Summing these conditions we obtain that $|(g(\mathbf{x}) - h(\mathbf{x})) - (g_T(\mathbf{x}) - h_T(\mathbf{x}))| \leqslant \varepsilon$, for all $\mathbf{x} \in \mathcal{K}$. The claim then immediately follows by recalling that $g(\mathbf{x}) - h(\mathbf{x}) = \phi(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{K}$, and letting $f_T \doteq g_T - h_T$, whence $f_T \in \mathrm{DLSE}_T$. $\square$

The following explicit example illustrates the approximation of a non-convex and nondifferentiable function by a function in LSE_T.

Example 1. Consider

$$\phi(x) = \max(0, \min(x, 1)).$$

Observe that $\phi(x) = \max(0, x) - \max(0, x - 1)$, which is indeed a difference of two nonsmooth convex functions. By using (5), we can approximate each term of this difference by a function in LSE as

$$\max(0, x) \leqslant T \log(1 + \exp(x/T))$$
$$\leqslant T \log 2 + \max(0, x),$$
$$\max(0, x - 1) \leqslant T \log(1 + \exp((x - 1)/T))$$
$$\leqslant T \log 2 + \max(0, x - 1).$$

It follows that the map

$$f_T \doteq T(\log(1 + \exp(x/T)) - \log\left(1 + \exp((x - 1)/T)\right))$$

is in DLSE_T and satisfies the following uniform approximation property of ϕ:

$$-T \log 2 + f_T(x) \leqslant \phi(x) \leqslant T \log 2 + f_T(x), \qquad \forall x \in \mathbb{R}^n.$$

Example 2. The previous explicit approximation carries over to a continuous piecewise affine function ϕ of a single real variable, as follows. By *piecewise affine*, we mean that $\mathbb{R}$ can be covered by finitely many intervals in such a way that ϕ is affine over each of these intervals. Then, ϕ can be written in a unique way as

$$\phi(x) = ax + b + \sum_{1 \leqslant i \leqslant K} \alpha_i \max(0, x - \gamma_i) \qquad (10)$$

where a, b are real parameters, $\gamma_1 < \cdots < \gamma_K$ are the nondifferentiability points of ϕ, and $\alpha_i = \phi'(\gamma_i^+) - \phi'(\gamma_i^-)$ is the jump of the derivative of ϕ at point γ_i. Another way to get insight of (10) is to make the following observation: the function ϕ has a second derivative in the distribution sense, $\phi'' = \sum_{1 \leqslant i \leqslant K} \alpha_i \delta_{\gamma_i}$; then (10) is gotten by integrating twice the latter expression of ϕ''. Possibly after subtracting to ϕ an affine function, we will always assume that $a = b = 0$.

Then, setting

$$I^+ \doteq \{1 \leqslant i \leqslant k \mid \alpha_i > 0\},$$
$$I^- \doteq \{1 \leqslant i \leqslant k \mid \alpha_i < 0\},$$

and

$$\phi^{\pm}(x) = \sum_{i \in I^{\pm}} |\alpha_i| \max(0, x - \gamma_i)$$

we write

$$\phi(x) = \phi^+(x) - \phi^-(x).$$

Let

$$a_i = \sum_{j \in I^{\pm}, j \leqslant i} |\alpha_j|,$$
$$b_i = \sum_{j \in I^{\pm}, j \leqslant i} |\alpha_j| \gamma_j,$$

and note that

$$\phi^{\pm}(x) = \max_{i \in I^{\pm}} (a_i x - b_i).$$

Then, setting $f_T \doteq f_T^+ - f_T^-$ where

$$f_T^{\pm}(x) = T \log \left(\sum_{i \in I^{\pm}} \exp((a_i x - b_i)/T) \right)$$

and using (5), we get

$$f_T - T \log |I^-| \leqslant f \leqslant f_T + T \log |I^+|.$$

C. Data approximation

Consider a collection $\mathcal{D}$ of m data-points,

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \qquad (11)$$

where $y_i = \phi(\mathbf{x}_i)$, $i = 1, \ldots, m$, and ϕ is an unknown function. The following universal data approximation result holds.

Corollary 2 (Universal data approximation)**.** *Given a collection of data $\mathcal{D}$ as in* (11)*, for any $\varepsilon > 0$ there exists $T > 0$ and a function $d_T \in \mathrm{DLSE}_T$ with rational coefficients such that*

$$|d_T(\mathbf{x}_i) - y_i| \leqslant \varepsilon, \quad i = 1, \ldots, m.$$

Proof. Let $\mathcal{K} \doteq \mathrm{co}\{\mathbf{x}_i, \ldots, \mathbf{x}_m\}$ be the convex hull of the input data points. Consider a triangulation of the input points $\mathbf{x}_i$: recall that such a triangulation consists of a finite collection of simplices $(\Delta_r)_{r \in R}$, satisfying the following properties: (i) the vertices of these simplices are taken among the points $\mathbf{x}_1, \ldots, \mathbf{x}_m$; (ii) each point $\mathbf{x}_i$ is the vertex of at least one simplex; (iii) the interiors of theses simplices have pairwise empty intersections, and (iv) the union of these simplices is precisely $\mathcal{K}$. Then, there is a unique continuous function, f, affine on each simplex Δ_r, and such that $f(\mathbf{x}_i) = y_i$ for $1 \leqslant i \leqslant m$. Observe that $\mathcal{K}$ is convex and compact by construction. Now, a direct application of Theorem 2 shows that for any $\varepsilon > 0$ there exists $T > 0$ and a function $d_T \in \mathrm{DLSE}_T$ with rational coefficients such that

$$|d_T(\mathbf{x}_i) - f(\mathbf{x}_i)| = |d_T(\mathbf{x}_i) - y_i| \leqslant \varepsilon, \quad i = 1, \ldots, m,$$

which concludes the proof. $\square$

IV. POSITIVE FUNCTIONS ON THE POSITIVE ORTHANT

In this section we discuss approximation results for functions taking positive values on the open positive orthant. A particular case of this class of functions is given by log-log-convex functions, whose uniform approximation by means of GPOS_T functions was discussed in Corollary 1 of [7]. We shall first extend this result to functions with rational parameters, and then provide a universal approximation result for continuous positive functions over the open positive orthant.

A. Uniform approximation results

The following preliminary definitions are instrumental for our purposes: a subset $\mathcal{R} \subset \mathbb{R}_{>0}^n$ will be said to be *log-convex* if its image by the map that takes the logarithm entry-wise is convex. We shall say that a function $\psi_T \in \mathrm{GPOS}_T$ has rational parameters if it can be written as in (7) with ψ given by (6), in such a way that T, the entries of the vectors $\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(K)}$, and the scalars $\log c_1, \ldots, \log c_K$ are rational numbers. The following corollary extends Corollary 1 of [7].

Corollary 3 (Universal approximators of log-log-convex functions)**.** *Let ℓ be a log-log-convex function defined on a compact, log-convex subset $\mathcal{R}$ of $\mathbb{R}_{>0}^n$. Then, for all $\tilde{\varepsilon} > 0$ there exist a function $\psi_T \in \mathrm{GPOS}_T$ with rational parameters, for some $T = 1/p$ where p is a positive integer, such that, for all $\mathbf{x} \in \mathcal{R}$,*

$$\left| \frac{\ell(\mathbf{x}) - \psi_T(\mathbf{x})}{\min(\ell(\mathbf{x}), \psi_T(\mathbf{x}))} \right| \leqslant \tilde{\varepsilon}. \qquad (12)$$

Proof. By using the log-log transformation, define $\tilde{\ell}(\mathbf{q}) \doteq \log(\ell(\exp(\mathbf{q})))$. Since $\ell(\mathbf{x})$ is log-log-convex in $\mathbf{x}$, $\tilde{\ell}(\mathbf{q})$ is convex in $\mathbf{q} = \log \mathbf{x}$. Furthermore, the set $\mathcal{K} \doteq \log(\mathcal{R})$ is convex and compact since the set $\mathcal{R}$ is log-convex and compact. Thus, by Corollary 1, for all $\varepsilon > 0$, there exist $T = 1/p$ where p is a positive integer, and a function $f_T \in \mathrm{LSE}_T$ with rational coefficients such that $|f_T(\mathbf{q}) - \tilde{\ell}(\mathbf{q})| \leqslant \varepsilon$ for all $\mathbf{q} \in \mathcal{K}$. From this point on, the proof follows the very same lines as the proof of Corollary 1 of [7]. $\square$

We next state an approximation result for functions on the positive orthant. The derivation of Theorem 3 from Corollary 3

is similar to the derivation of Theorem 2 from Corollary 1 and thus we omit its proof.

Theorem 3 (Universal approximators of functions on the open orthant). *Let ℓ be a continous positive function defined on a compact log-convex subset $\mathcal{R} \subset \mathbb{R}^n_{>0}$. Then, for all $\tilde{\varepsilon} > 0$ there exist two functions $\psi_T, \psi'_T \in \mathrm{GPOS}_T$ with rational parameters, for some $T = 1/p$ where p is a positive integer, such that, for all $\mathbf{x} \in \mathcal{R}$,*

$$\left| \frac{\ell(\mathbf{x}) - \psi_T(\mathbf{x})/\psi'_T(\mathbf{x})}{\min(\ell(\mathbf{x}), \psi_T(\mathbf{x})/\psi'_T(\mathbf{x}))} \right| \leqslant \tilde{\varepsilon}. \tag{13}$$

B. Universal approximation by subtraction-free expressions

We next derive from Theorem 3 an approximation result by *subtraction-free expressions*. The latter are an important subclass of rational expressions, studied in [8]. Subtraction-free expressions are well formed expressions in several commutative variables $x_1, \ldots, x_n$, defined using the operations $+, \times, /$ and using positive constants, but not using subtraction. Formally, a subtraction-free expression in the variables $x_1, \ldots, x_n$ is a term produced by the context-free grammar rule

$$E \to E + E, E \times E, E/E, C, x_1, \ldots, x_n$$

where C can take the value of any positive constant. For instance, $E_1 \doteq (x_1 + x_2^3)/(2x_1 + 3x_2/(x_1 + x_2))$ is a subtraction-free expression, whereas $E_2 \doteq x_1^2 - x_1 x_2 + x_2^2$ is not a subtraction-free expression, owing to the presence of the $-$ sign. Note that E_2, thought of as a formal rational fraction, coincides with $E_3 \doteq (x_1^3 + x_2^3)/(x_1 + x_2)$ which is subtraction free, i.e., an expression which is not subtraction-free may well have an equivalent subtraction-free expression. However, there are rational fractions, and even polynomials, like $E_4 \doteq (x_1 - x_2)^2$, without subtraction equivalent free expressions, because any subtraction-free expression must take positive values on the interior of the positive cone, whereas E_4 vanishes on the line $x_1 = x_2$. Important examples of subtraction-free expressions arise from series-parallel composition rules for resistances. More advanced examples, coming from algebraic combinatorics, are discussed in [8].

Corollary 4 (Approximation by subtraction-free expressions). *Let ℓ be a continous positive function defined on a compact log-convex subset $\mathcal{R} \subset \mathbb{R}^n_{>0}$. Then, for all $\tilde{\varepsilon} > 0$ there exist positive integers p, q and a subtraction-free expression E in n variables $y_1, \ldots, y_n$ such that the function*

$$f(\mathbf{x}) = E(x_1^{1/q}, \ldots, x_n^{1/q})^{1/p}$$

in which $x_i^{1/q}$ is substituted to the variable y_i, satisfies, for all $\mathbf{x} \in \mathcal{R}$,

$$\left| \frac{\ell(\mathbf{x}) - f(\mathbf{x})}{\min(\ell(\mathbf{x}), f(\mathbf{x}))} \right| \leqslant \tilde{\varepsilon}. \tag{14}$$

Proof. Theorem 3 shows that (14) holds with $f = \psi_T/\psi'_T$ where $T = 1/p$ for some positive integer p, and ψ_T, ψ'_T are functions in GPOS_T with rational parameters, i.e.,

$$\psi_T(\mathbf{x}) = \left(\sum_{k=1}^K c_k \mathbf{x}^{\boldsymbol{\alpha}^{(k)}p} \right)^{1/p},$$

$$\psi'_T(\mathbf{x}) = \left(\sum_{k=1}^{K'} c'_k \mathbf{x}^{(\boldsymbol{\alpha}')^{(k)}p} \right)^{1/p},$$

where the vectors $\boldsymbol{\alpha}^{(k)}$ and $(\boldsymbol{\alpha}')^{(k)}$ have rational entries. Denoting by q the least common multiple of the denominators of the entries of the vectors $\boldsymbol{\alpha}^{(k)}p$ and $(\boldsymbol{\alpha}')^{(k)}p$, we see that $\psi_T(\mathbf{x})/\psi_T(\mathbf{x})$ is precisely of the form $E(x_1^{1/q}, \ldots, x_n^{1/q})^{1/p}$ where E is a subtraction-free rational expression. $\square$

C. Approximation of positive data

Consider a collection $\mathcal{L}$ of m data pairs,

$$\mathcal{L} = \{(\mathbf{z}_i, w_i)\}_{i=1}^m,$$

where $\mathbf{z}_i \in \mathbb{R}^n_{>0}$, $w_i \in \mathbb{R}_{>0}$, $i = 1, \ldots, m$, with $w_i = \ell(\mathbf{z}_i)$, $i = 1, \ldots, m$, where $\ell : \mathbb{R}^n_{>0} \to \mathbb{R}_{>0}$ is an unknown function. The data in $\mathcal{L}$ is referred to as *positive data*. The following proposition is now an immediate consequence of Theorem 3, where $\mathcal{R}$ can be taken as the log-convex hull of the input data points[2].

Proposition 1. *Given positive data $\mathcal{L} \doteq \{(\mathbf{z}_i, w_i)\}_{i=1}^m$, for any $\tilde{\varepsilon} > 0$ there exist a rational $T > 0$ and two functions $\psi_T, \psi'_T \in \mathrm{GPOS}_T$ with rational parameters such that*

$$\left| \frac{w_i - \psi_T(\mathbf{z}_i)/\psi'_T(\mathbf{z}_i)}{\min(w_i, \psi_T(\mathbf{z}_i)/\psi'_T\mathbf{z}_i))} \right| \leqslant \tilde{\varepsilon}, \quad i = 1, \ldots, m. \tag{15}$$

V. DLSE_T NETWORKS

Functions in DLSE_T can be modeled through a feedforward neural network (FFNN) architecture, composed of two LSE_T networks in parallel, whose outputs are fused via an output difference node, see Fig. 1.

It may sometimes be convenient to highlight the full parameterization of the input-output function $d_T(\cdot)$ synthesized by the DLSE_T network, in which case we shall write

$$d_T^{(\overrightarrow{\boldsymbol{\alpha}}, \overrightarrow{\boldsymbol{\gamma}}, \boldsymbol{\beta}, \boldsymbol{\delta})}(\mathbf{x}) = f_T^{(\overrightarrow{\boldsymbol{\alpha}}, \boldsymbol{\beta})}(\mathbf{x}) - f_T^{(\overrightarrow{\boldsymbol{\gamma}}, \boldsymbol{\delta})}(\mathbf{x})$$

where $\overrightarrow{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(K)})$, $\overrightarrow{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(K)})$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$, and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_k)$ are the parameter vectors of the two LSE_T components.

Each LSE_T component has n input nodes, one hidden layer with K nodes, and one output node. The activation function of the hidden nodes is $s \mapsto (\exp(s/T))$, and the activation of the output node of each LSE_T component is $s \mapsto T \log(s)$. Each node in the hidden layer of the first LSE_T component network computes a term of the form $s_k = \langle \boldsymbol{\alpha}^{(k)}, \mathbf{x} \rangle + \beta_k$, where the i-th entry $\alpha_i^{(k)}$ of $\boldsymbol{\alpha}^{(k)}$ represents the weight between node k

---

[2]For given $\mathbf{z}_1, \ldots, \mathbf{z}_m \in \mathbb{R}^n_{>0}$, we define their log-convex hull as the set of vectors $\mathbf{z} = \prod_{i=1}^m \mathbf{z}_i^{\xi_i}$, where $\xi_i \in [0, 1]$, $i = 1, \ldots, m$, and $\sum_{i=1}^m \xi_i = 1$.

Fig. 1. A DLSE$_T$ network is composed of two LSE$_T$ networks in parallel, with a difference output node.

Fig. 2. The same DLSE$_T$ network as in Fig. 1 can be obtained by suitably pre-scaling the input and outputs of the two component LSE networks.

and input x_i, and β_k is the bias term of node k. Each node k of the first LSE$_T$ network thus generates activations

$$a_k = \exp(\langle \boldsymbol{\alpha}^{(k)}, \mathbf{x}/T \rangle + \beta_k/T).$$

We consider the weights from the inner nodes to the output node to be unitary, whence the output node of the first LSE$_T$ network computes $s = \sum_{k=1}^{K} a_k$ and then, according to the output activation function, the output layer returns the value

$$T \log(s) = T \log \left(\sum_{k=1}^{K} a_k \right).$$

An identical reasoning applies to the output of the second component LSE$_T$ network. The overall output realizes a DLSE$_T$ function which, by Theorem 2, allows us to approximate any continuous function over a compact convex domain. Similarly, by Corollary 2, we can approximate data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$ via a DLSE$_T$ network, to any given precision.

Theorem 4. *Given a collection of data $\mathcal{D} \doteq \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, for each $\varepsilon > 0$ there exists a DLSE$_T$ neural network such that*

$$|d_T(\mathbf{x}_i) - y_i| \leqslant \varepsilon, \quad i = 1, \ldots, m,$$

where d_T is the input-output function of the network.

A. Training DLSE$_T$ networks

By using the scaling property (4) of LSE$_T$ functions, it can be noticed that a simpler LSE network structure can be used to implement a DLSE$_T$ neural network, as shown in Fig. 2.

As a matter of fact, given the parameter vectors $\overrightarrow{\boldsymbol{\alpha}}$, $\overrightarrow{\boldsymbol{\gamma}}$, $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, it can be easily derived that

$$
\begin{aligned}
d_T^{(\overrightarrow{\boldsymbol{\alpha}}, \overrightarrow{\boldsymbol{\gamma}}, \boldsymbol{\beta}, \boldsymbol{\delta})}(\mathbf{x}) &= f_T^{(\overrightarrow{\boldsymbol{\alpha}}, \boldsymbol{\beta})}(\mathbf{x}) - f_T^{(\overrightarrow{\boldsymbol{\gamma}}, \boldsymbol{\delta})}(\mathbf{x}) \\
&= T \left(f_1^{(\overrightarrow{\boldsymbol{\alpha}}, \boldsymbol{\beta}/T)}(\mathbf{x}/T) - f_1^{(\overrightarrow{\boldsymbol{\gamma}}, \boldsymbol{\delta}/T)}(\mathbf{x}/T) \right) \\
&= T d_1^{(\overrightarrow{\boldsymbol{\alpha}}, \overrightarrow{\boldsymbol{\gamma}}, \boldsymbol{\beta}/T, \boldsymbol{\delta}/T)}(\mathbf{x}/T),
\end{aligned}
$$

and hence dealing with a DLSE$_T$ neural network corresponds to deal with a DLSE neural network whose input and output have been rescaled. Each of the two LSE components of the network realizes an input-output map of the form

$$f^{(\overrightarrow{\boldsymbol{\alpha}}, \boldsymbol{\beta}/T)}(\mathbf{x}/T) = \log \left(\sum_{k=1}^{K} \exp(\langle \boldsymbol{\alpha}^{(k)}, \mathbf{x}/T \rangle + \beta_k/T) \right).$$

The gradients of this function with respect to its parameters $\boldsymbol{\alpha}^{(i)}$, β_i are, for $i = 1, \ldots, K$,

$$\nabla_{\boldsymbol{\alpha}^{(i)}} f^{(\overrightarrow{\boldsymbol{\alpha}}, \boldsymbol{\beta}/T)}(\mathbf{x}/T) = \frac{\exp(\langle \boldsymbol{\alpha}^{(i)}, \mathbf{x}/T \rangle + \beta_i/T)\, \mathbf{x}}{T \sum_{k=1}^{K} \exp(\langle \boldsymbol{\alpha}^{(k)}, \mathbf{x}/T \rangle + \beta_k/T)},$$

$$\nabla_{\beta_i} f^{(\overrightarrow{\boldsymbol{\alpha}}, \boldsymbol{\beta}/T)}(\mathbf{x}/T) = \frac{\exp(\langle \boldsymbol{\alpha}^{(i)}, \mathbf{x}/T \rangle + \beta_i/T)}{T \sum_{k=1}^{K} \exp(\langle \boldsymbol{\alpha}^{(k)}, \mathbf{x}/T \rangle + \beta_k/T)}.$$

Thus, by using the chain rule, we have that

$$\nabla_{\boldsymbol{\alpha}^{(i)}} d_T^{(\overrightarrow{\boldsymbol{\alpha}}, \overrightarrow{\boldsymbol{\gamma}}, \boldsymbol{\beta}, \boldsymbol{\delta})}(\mathbf{x}) = \frac{\exp(\langle \boldsymbol{\alpha}^{(i)}, \mathbf{x}/T \rangle + \beta_i/T)\, \mathbf{x}}{\sum_{k=1}^{K} \exp(\langle \boldsymbol{\alpha}^{(k)}, \mathbf{x}/T \rangle + \beta_k/T)},$$

$$\nabla_{\beta_i} d_T^{(\overrightarrow{\boldsymbol{\alpha}}, \overrightarrow{\boldsymbol{\gamma}}, \boldsymbol{\beta}, \boldsymbol{\delta})}(\mathbf{x}) = \frac{\exp(\langle \boldsymbol{\alpha}^{(i)}, \mathbf{x}/T \rangle + \beta_i/T)}{\sum_{k=1}^{K} \exp(\langle \boldsymbol{\alpha}^{(k)}, \mathbf{x}/T \rangle + \beta_k/T)},$$

$$\nabla_{\boldsymbol{\gamma}^{(i)}} d_T^{(\overrightarrow{\boldsymbol{\alpha}}, \overrightarrow{\boldsymbol{\gamma}}, \boldsymbol{\beta}, \boldsymbol{\delta})}(\mathbf{x}) = -\frac{\exp(\langle \boldsymbol{\gamma}^{(i)}, \mathbf{x}/T \rangle + \delta_i/T)\, \mathbf{x}}{\sum_{k=1}^{K} \exp(\langle \boldsymbol{\gamma}^{(k)}, \mathbf{x}/T \rangle + \delta_k/T)},$$

$$\nabla_{\delta_i} d_T^{(\overrightarrow{\boldsymbol{\alpha}}, \overrightarrow{\boldsymbol{\gamma}}, \boldsymbol{\beta}, \boldsymbol{\delta})}(\mathbf{x}) = -\frac{\exp(\langle \boldsymbol{\gamma}^{(i)}, \mathbf{x}/T \rangle + \delta_i/T)}{\sum_{k=1}^{K} \exp(\langle \boldsymbol{\gamma}^{(k)}, \mathbf{x}/T \rangle + \delta_k/T)}.$$

Given a dataset $\mathcal{D}$ as in (11), these gradients can be used to train a DLSE$_T$ network by using classical algorithms

such as the Levenberg-Marquardt algorithm [21], the Fletcher-Powell conjugate gradient [22], or the stochastic gradient descent algorithm [23]. In numerical practice, one may fix the parameter T and K and train the network with respect to the parameters $\overrightarrow{\boldsymbol{\alpha}}$, $\boldsymbol{\beta}$, $\overrightarrow{\boldsymbol{\gamma}}$ and $\boldsymbol{\delta}$ by using one of the methods mentioned above, until a satisfactory cross-validated fit is found. A suitable initial value for the T parameter may be set, for instance, to the inverse mid output range $2/|\max(y_i) - \min(y_i)|$. Alternatively, T can be considered as a trainable variable as well, and computed by the training algorithm alongside with the other parameters.

The following example illustrates the application of the Levenberg-Marquardt algorithm to a simple case.

Example 3. Consider the function $\phi : [-2, 2] \to \mathbb{R}$,

$$\phi(x) = x^2 + \sin(2\pi x). \quad (16)$$

Such a function, which is clearly nonconvex, has been used to generate the dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{100}$, where each x_i has been taken uniformly at random in $[-2, 2]$ and $y_i = \phi(x_i)$, $i = 1, \ldots, 100$. The Levenberg-Marquardt algorithm has been used to train a DLSE_T network fitting such data with $K = 10$ and $T = 2/|\max(y_i) - \min(y_i)|$. Fig. 3 depicts the output of the DLSE_T network and of its two LSE_T components $f_T^{(\overrightarrow{\boldsymbol{\alpha}},\boldsymbol{\beta})}(\mathbf{x})$ and $f_T^{(\overrightarrow{\boldsymbol{\gamma}},\boldsymbol{\delta})}(\mathbf{x})$.

Fig. 3. Output of the trained DLSE_T network.

As shown in Fig. 3, although the function ϕ is nonconvex, it is well approximated by a DLSE_T network. Indeed, the data represented in Fig. 3 are approximated by the trained DLSE_T network with a mean square error of $4.4 \cdot 10^{-5}$.

VI. NON-CONVEX OPTIMIZATION VIA DLSE_T NETWORKS

In view of the results established in Sections III and V, DLSE_T networks can be efficiently used to compute a *difference of convex* (*DC*) approximate decomposition of any continuous function over a compact set. Indeed, by using the tools described in Section V, given any continuous function

$\phi(\cdot)$ defined on a convex compact set $\mathcal{K} \subset \mathbb{R}^n$ (or, more generally, a dataset generated through any function ϕ) and $\varepsilon \in \mathbb{R}_{>0}$, we can determine $g_T, h_T \in \text{LSE}_T$ such that

$$|\phi(\mathbf{x}) - g_T(\mathbf{x}) + h_T(\mathbf{x})| \leqslant \varepsilon, \quad \forall \mathbf{x} \in \mathcal{K}. \quad (17)$$

Once functions $g_T(\mathbf{x})$ and $h_T(\mathbf{x})$ are determined via training on available data, we have a surrogate model $d_T = g_T - h_T \simeq \phi$ that we can use for solving approximately design problems of the form $\min_{\mathbf{x} \in \mathcal{K}} \phi(\mathbf{x})$, by substituting the surrogate function d_T in place of the (possibly unknown) function ϕ. The resulting surrogate design problem

$$\min_{\mathbf{x} \in \mathcal{K}} d_T^{(\overrightarrow{\boldsymbol{\alpha}}, \overrightarrow{\boldsymbol{\gamma}}, \boldsymbol{\beta}, \boldsymbol{\delta})}(\mathbf{x}) \quad (18)$$

involves the minimization of the difference of two convex LSE_T functions. An approximate solution to this problem can be computed by means of a specific and effective algorithm named Difference-of-Convex Algorithm (DCA), which is described in [24], [25], [26], [27]. We next tailor the DCA to our specific context in the following Algorithm 1, denoted by DLSEA.

Algorithm 1 Difference of LSE_T algorithm (DLSEA)

Input: functions $g_T = f_T^{(\overrightarrow{\boldsymbol{\alpha}}, \boldsymbol{\beta})}$ and $h_T = f_T^{(\overrightarrow{\boldsymbol{\gamma}}, \boldsymbol{\delta})}$ in LSE_T and a convex compact set $\mathcal{K}$.

Output: a candidate optimal solution $\hat{\mathbf{x}}^\star$ to the problem

$$\min_{\mathbf{x} \in \mathcal{K}} (g_T(\mathbf{x}) - h_T(\mathbf{x})). \quad (19)$$

1: pick initial point $\boldsymbol{\chi}^{(0)} \in \mathcal{K}$
2: **for** $\varkappa \in \mathbb{N}$ **do**
3:     let $\mathbf{v}^{(\varkappa)} = \dfrac{\sum_{k=1}^{K} \exp(\langle \boldsymbol{\gamma}^{(k)}, \boldsymbol{\chi}^{(\varkappa)}/T \rangle + \delta_k/T) \, \boldsymbol{\gamma}^{(k)}}{\sum_{k=1}^{K} \exp(\langle \boldsymbol{\gamma}^{(k)}, \boldsymbol{\chi}^{(\varkappa)}/T \rangle + \delta_k/T)}$
4:     let $\boldsymbol{\chi}^{(\varkappa+1)} = \arg\min_{\mathbf{x} \in \mathcal{K}} \{g_T(\mathbf{x}) - \langle \mathbf{x}, \mathbf{v}^{(\varkappa)} \rangle\}$
5:     **if** $\dfrac{\|\boldsymbol{\chi}^{(\varkappa+1)} - \boldsymbol{\chi}^{(\varkappa)}\|}{1 + \|\boldsymbol{\chi}^{(\varkappa)}\|}$ is smaller than a tolerance **then**
6:         **return** $\hat{\mathbf{x}}^\star = \boldsymbol{\chi}^{(\varkappa+1)}$

The DLSEA returns a candidate optimal solution to problem (19), see [27]. In words, this algorithm starts from a given initial point $\boldsymbol{\chi}^{(0)} \in \mathcal{K}$ and iterates until convergence Step 3, in which the gradient of $h_T = f_T^{(\overrightarrow{\boldsymbol{\gamma}}, \boldsymbol{\delta})}$ is computed as $\mathbf{v}^{(\varkappa)} = \nabla h_T(\boldsymbol{\chi}^{(\varkappa)})$, and Step 4, in which a local approximation of problem (19) is solved. We observe that the function $g_T(\mathbf{x}) - \langle \mathbf{x}, \mathbf{v}^{(\varkappa)} \rangle$ minimized in Step 4 is (except for a constant term that does not affect the minimization) equal to the difference between the convex function $g_T(\mathbf{x})$ and the linearization of h_T around the current solution $\boldsymbol{\chi}^{(\varkappa)}$. Therefore, the problem to be solved in Step 4 is a convex minimization problem, which can be solved globally and efficiently via standard tools.

Example 4. The DLSEA has been used to find the global minimum of the function ϕ given in (16) on the compact convex set $\mathcal{K} = [-2, 2]$. Namely, by exploiting the approximate DC decomposition of ϕ determined in Example 3 and letting $\boldsymbol{\chi}^{(0)} = 0$, we obtained, with 4 iterations of the DLSEA, the approximate solution $\hat{\mathbf{x}}^\star = -0.2381$ to problem (19), that is close to the actual solution $\mathbf{x}^\star = -0.2379$.

VII. Application: diet design for type 2 diabetes

Type 2 diabetes mellitus is a chronic disease that affects the way the human body processes glucose. It is characterized by reduced sensitivity of tissues to insulin, a hormone produced by pancreatic beta cells that promotes the absorption of glucose from blood into liver, fat and skeletal muscle cells [28].

The main objective of this section is to show how DLSE_T networks and the DLSEA can be used to design a diet based on 5 meals for a patient with type 2 diabetes, with the aim of minimizing the maximal concentration of glucose in blood, while guaranteeing that a sufficient amount of glucose is administrated (namely, exactly $185\,\text{g}$). In order to pursue this objective, the meal model for the glucose-insulin system given in [29] has been used to simulate the time-behavior of plasma glucose concentration with breakfast at 8 a.m. (containing $x_1\,\text{g}$ of glucose), mid-morning snack at 11 a.m. (containing $x_2\,\text{g}$ of glucose), lunch at 1 p.m. (containing $x_3\,\text{g}$ of glucose), mid-afternoon snack at 5 p.m. (containing $x_4\,\text{g}$ of glucose), and dinner at 8 p.m. (containing $x_5\,\text{g}$ of glucose). This model has been used to generate synthetic data: 10^3 points $\mathbf{x}^{(i)} = [\; x_1^{(i)} \quad \cdots \quad x_5^{(i)} \;]^\top$ have been picked uniformly at random from the set

$$\mathcal{K} = \left\{ \mathbf{x} \in \mathbb{R}^5 : x_j \geqslant 0,\, j = 1, \ldots, 5,\, \sum_{j=1}^{5} x_j = 185 \right\},$$

and the meal model has been used to determine the corresponding maximum of glucose concentration $y^{(i)}$. Thus, the training dataset $\mathcal{D}_{\text{train}} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{1000}$ has been used to train a DLSE_T network with temperature parameter $T = 12.98 \cdot 10^{-3}$ (that is $2/|\max(y_i) - \min(y_i)|$) and with $K = 30$ nodes in each of its two LSE_T components. In order to evaluate the prediction capabilities of this network, after the training, we generated, by using the same method as above, a validation dataset $\mathcal{D}_{\text{valid}} = \{(\check{\mathbf{x}}^{(i)}, \check{y}^{(i)})\}_{i=1}^{100}$, and we compared the outputs of the DLSE_T model $d_T(\check{\mathbf{x}}^{(i)})$ with $\check{y}^{(i)}$, $i = 1, \ldots, 100$. For comparison purposes, a classical FFNN with symmetric sigmoid activation function for the hidden layer (with 60 nodes) and linear activation function for the output layer has been trained on the same dataset $\mathcal{D}_{\text{train}}$ and its prediction performance has been evaluated over $\mathcal{D}_{\text{valid}}$. Table I summarizes the prediction errors of the two models.

TABLE I
PREDICTION ERRORS OVER $\mathcal{D}_{\text{valid}}$

Method	Mean Sq. [mg/dL]	Mean Rel. [−]	Max Abs. [mg/dL]	Max Rel. [−]	r^2 [−]
DLSE	0.0624	0.0006	1.1043	0.0043	0.9999
FFNN	2.3145	0.0041	7.9480	0.0311	0.9959

As shown by Table I, the DLSE_T model has the best performance with respect to all the error metrics. Furthermore, differently form classical FFNN, it is readily amenable to efficient optimization via the DLSEA. Indeed, we applied such an algorithm to the DLSE_T model and we obtained the optimal diet that minimizes the maximal concentration of glucose in blood, while guaranteeing that $185\,\text{g}$ of glucose are administrated. Fig. 4 depicts the optimal diet and the

corresponding time-behavior of the plasma glucose concentration. The optimal diet is such that the corresponding maximal plasma glucose concentration in $24\,\text{h}$ is $253.06\,\text{mg/dL}$.

Fig. 4. Simulation of the time-behavior of the plasma glucose concentration with the optimal diet determined via the DLSEA.

VIII. Conclusions

In this paper, we showed that a neural network whose output is the difference of the outputs of two feedforward neural networks with exponential activation function in the hidden layer and logarithmic activation function in the output node is an universal approximator of continuous functions over compact convex sets. By using a logarithmic transformation, such networks maps to a class of subtraction free ratios of generalized posynomials, which we showed to be universal approximators of positive functions over the positive orthant.

The main advantage of DLSE_T networks with respect to classical FFNN is that they are readily amenable to effective optimization-based design. In particular, by adapting the DCA given in [27] to our context, we derived an ad-hoc algorithm for optimizing DLSE_T models which has proved to be efficient in the considered test cases.

References

[1] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control Signals Syst.*, vol. 2, no. 4, pp. 303–314, 1989.

[2] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[3] A. H. Zaabab, Q.-J. Zhang, and M. Nakhla, "A neural network modeling approach to circuit optimization and statistical design," *IEEE Trans. Microwave Theory Tech.*, vol. 43, no. 6, pp. 1349–1358, 1995.

[4] J. G. Kuschewski, S. Hui, and S. H. Zak, "Application of feedforward neural networks to dynamical system identification and control," *IEEE Trans. Control Syst. Technol.*, vol. 1, no. 1, pp. 37–49, 1993.

[5] C. Peterson and B. Söderberg, "A new method for mapping optimization problems onto neural networks," *Int. J. Neural Syst.*, vol. 1, no. 01, pp. 3–22, 1989.

[6] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[7] G. Calafiore, S. Gaubert, and C. Possieri, "Log-sum-exp neural networks and posynomial models for convex and log-log-convex data," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[8] S. Fomin, D. Grigoriev, and G. Koshevoy, "Subtraction-free complexity, cluster transformations, and spanning trees," *Found. Comput. Math.*, vol. 16, pp. 1–31, Feb. 2016.

[9] O. Viro, "Dequantization of real algebraic geometry on logarithmic paper," in *European Congress of Mathematics, Vol. I (Barcelona, 2000)*, vol. 201 of *Progr. Math.*, pp. 135–146, Birkhäuser, Basel, 2001.

[10] I. Itenberg, G. Mikhalkin, and E. Shustin, *Tropical algebraic geometry*. Oberwolfach seminars, Birkhäuser, 2007.

[11] V. Charisopoulos and P. Maragos, "Morphological perceptrons: Geometry and training algorithms," in *Mathematical Morphology and Its Applications to Signal and Image Processing* (J. Angulo, S. Velasco-Forero, and F. Meyer, eds.), pp. 3–15, Springer Int. Publishing, 2017.

[12] L. Zhang, G. Naitzat, and L.-H. Lim, "Tropical geometry of deep neural networks," in *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018*, 2018. To appear.

[13] S. Ovchinnikov, "Max-min representations of piecewise linear functions," *Beitr. Algebra Geom.*, vol. 43, no. 1, pp. 297–302, 2002.

[14] S. Wang, "General constructive representations for continuous piecewise-linear functions," *IEEE Transactions on Circuits and Systems – I: Regular Papers*, vol. 51, no. 9, pp. 1889–1896, 2004.

[15] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pp. III–1319–III–1327, JMLR.org, 2013.

[16] Y. Zhang, S. Blusseau, S. Velasco-Forero, I. Bloch, and J. Angulo, "Max-plus operators applied to filter selection and model pruning in neural networks," 2019.

[17] B. Klartag and E. Milman, "Centroid bodies and the logarithmic Laplace transform–a unified approach," *J. Funct. Anal.*, vol. 262, no. 1, pp. 10–34, 2012.

[18] M. G. Crandall and L. Tartar, "Some relations between nonexpansive and order preserving mappings," *Proc. Amer. Math. Soc.*, vol. 78, no. 3, pp. 385–390, 1980.

[19] M. Akian, S. Gaubert, and A. Hochart, "Minimax representation of nonexpansive functions and application to zero-sum recursive games," *J. Convex Anal.*, no. 1, pp. 225–240, 2018.

[20] M. Bačák and J. M. Borwein, "On difference convexity of locally Lipschitz functions," *Optimization*, vol. 60, no. 8-9, pp. 961–978, 2011.

[21] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *J. Soc. Ind. Appl. Math.*, vol. 11, no. 2, pp. 431–441, 1963.

[22] W. C. Davidon, "Variable metric method for minimization," *SIAM J. Optim.*, vol. 1, no. 1, pp. 1–17, 1991.

[23] D. P. Bertsekas, *Nonlinear programming*. Athena scientific, 1999.

[24] T. Pham Dinh, "The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems," *Ann. Oper. Res.*, vol. 133, no. 1-4, pp. 23–46, 2005.

[25] H. A. Le Thi, H. M. Le, and T. Pham Dinh, "A DC programming approach for feature selection in support vector machines learning," *Adv. Data Anal. Classification*, vol. 2, no. 3, pp. 259–278, 2008.

[26] T. Pham Dinh and H. A. Le Thi, "Recent advances in DC programming and DCA," in *Trans. Comput. Int. XIII*, pp. 1–37, Springer, 2014.

[27] H. A. Le Thi and T. Pham Dinh, "DC programming and DCA: thirty years of developments," *Math. Prog.*, vol. 169, no. 1, pp. 5–68, 2018.

[28] C. M. Ripsin, H. Kang, and R. J. Urban, "Management of blood glucose in type 2 diabetes mellitus," *Am. Fam. Physician*, vol. 79, no. 1, pp. 29–36, 2009.

[29] C. Dalla Man, R. A. Rizza, and C. Cobelli, "Meal simulation model of the glucose-insulin system," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 10, pp. 1740–1749, 2007.

Giuseppe C. Calafiore (S '14, F '18) received the "Laurea" degree in Electrical Engineering from Politecnico di Torino in 1993, and the Ph.D. degree in Information and System Theory from Politecnico di Torino, in 1997. He is with the faculty of Dipartimento di Electronica e Telecommunicazioni, Politecnico di Torino, where he currently serves as a full professor and coordinator of the Systems and Data Science lab. He is associated with the Italian National Research Council (CNR). Dr. Calafiore held several visiting positions at international institutions: at the Information Systems Laboratory (ISL), Stanford University, California, in 1995; at the Ecole Nationale Supérieure de Techniques Avanceés (ENSTA), Paris, in 1998; and at the University of California at Berkeley, in 1999, 2003 and 2007. He had an appointment as a Senior Fellow at the Institute of Pure and Applied Mathematics (IPAM), University of California at Los Angeles, in 2010. He had appointments as a Visiting Professor at EECS UC Berkeley in 2017 and at the Haas Business School in 2018 and 2019. Dr. Calafiore is the author of more than 180 journal and conference proceedings papers, and of eight books. He is a fellow member of the IEEE since 2018. He received the IEEE Control System Society "George S. Axelby" Outstanding Paper Award in 2008. His research interests are in the fields of convex optimization, randomized algorithms, machine learning, computational finance, and identification and control of uncertain systems.

Stephane Gaubert (M '18) obtained the Engineer degree from École Polytechnique, Palaiseau, in 1988. He got a PhD degree in Mathematics and Automatic Control from École Nationale Supérieure des Mines de Paris in 1992. He is senior research scientist (Directeur de Recherche) at INRIA Saclay – Île-de-France and member of CMAP (Centre de Mathématiques Appliquées, École Polytechnique, CNRS), head of a joint research team, teaching at École Polytechnique. He coordinates the Gaspard Monge corporate sponsorship Program for Optimization and Operations Research (PGMO), of Fondation Mathématique Hadamard, Paris-Saclay. His interests include tropical geometry, optimization, game theory, monotone or nonexpansive dynamical systems, and applications of mathematics to decision making and to the verification of programs or systems.

Corrado Possieri received his bachelor's and master's degrees in Medical engineering and his Ph.D. degree in Computer Science, Control and Geoinformation from the University of Roma Tor Vergata, Italy, in 2011, 2013, and 2016, respectively. From September 2015 to June 2016, he visited the University of California, Santa Barbara (UCSB). Currently, he is Assistant Professor at the Politecnico di Torino. He is a member of the IFAC TC on Control Design. His research interests include stability and control of hybrid systems, the application of computational algebraic geometry techniques to control problems, stochastic systems, and optimization.