Convex Dual Theory Analysis of Two-Layer Convolutional Neural Networks With Soft-Thresholding

Chunyan Xiong, Chaoxing Zhang, Mengli Lu, Xiaotong Yu, Jian Cao, Zhong Chen, Di Guo, *Member, IEEE*, and Xiaobo Qu, *Senior Member, IEEE*

***Abstract*— Soft-thresholding has been widely used in neural networks. Its basic network structure is a two-layer convolution neural network with soft-thresholding. Due to the network's nature of nonlinear and nonconvex, the training process heavily depends on an appropriate initialization of network parameters, resulting in the difficulty of obtaining a globally optimal solution. To address this issue, a convex dual network is designed here. We theoretically analyze the network convexity and prove that the strong duality holds. Extensive results on both simulation and real-world datasets show that strong duality holds, the dual network does not depend on initialization and optimizer, and enables faster convergence than the state-of-the-art two-layer network. This work provides a new way to convexify soft-thresholding neural networks. Furthermore, the convex dual network model of a deep soft-thresholding network with a parallel structure is deduced.**

***Index Terms*— Convex optimization, nonconvexity, soft-thresholding, strong duality.**

I. Introduction

NEURAL networks (NNs) have been extensively employed in various applications, including speech and image recognition [1], [2], image classification [3], fast medical imaging [4], and biological spectrum reconstruction [5], [6], [7]. NN, however, is easy to stuck at the local optimum or the saddle point due to the network nonconvexity (Fig. 1) [8]. This limitation prevents NN from reaching the global optimum [9], [10], [11]. To address this issue, proper initialization of network parameters is required in the training process [12], [13], [14].

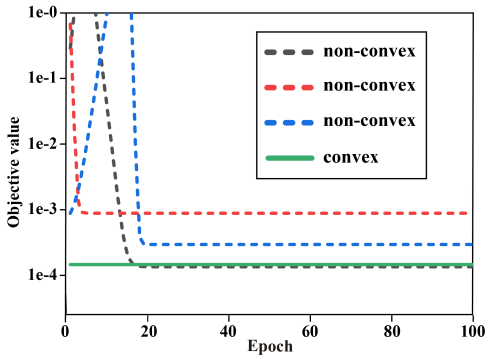


Fig. 1. Toy example: In some bad cases, the nonconvex NN gets stuck in a local optimum or saddle point. The objective value of a two-layer nonconvex and a convex NN model for 1-D vector fitting. The input of the network is $[-2, -1, 0, 1, 2]^\top$ and its ideal output (also called the label) is $[1, -1, -1, -1, 1]^\top$. Here, the bias term is included by concatenating a column of ones to the input. Under three random initialization trials of network parameters, the objective value of the nonconvex NN is different. Convex NNs, which do not depend on the initialization, can be solved directly with a convex procedure to obtain the optimal value.

Typical initialization strategies have been established [3], [15], [16] but the network may still encounter instability if the NN has multiple layers or branches [13]. For example, the original Transformer model [17] did not converge without initializing the learning rate in a warm-up way [18], [19], [20]. Roberta [21] and GPT-3 [22] had to tune the parameters of the optimizer ADAM [23] for stability under the large batch size. Recent studies have shown that architecture-specific initialization can promote convergence [19], [24], [25], [26], [27]. Even though, these initialization techniques hardly work to their advantage when conducting architecture searches, training networks with branching or heterogeneous components [13].

Convexifying NNs is another way to make the solution not depend on initialization [28], [29]. At present, theoretical research on the convexification of NN focuses on finite-width networks which include fully connected networks [30], [31], [32], [33], [34], [35], [36], [37] and convolutional NNs (CNNs) [38], [39]. The former is powerful to learn multilevel features [40] but may require a large space and computational

Manuscript received 11 April 2023; revised 4 November 2023; accepted 9 January 2024. This work was supported in part by the National Natural Science Foundation under Grant 61971361, Grant 62122064, Grant 62331021, and Grant 62371410; in part by the Natural Science Foundation of Fujian Province of China under Grant 2023J02005 and Grant 2021J011184; in part by the President Fund of Xiamen University under Grant 20720220063; and in part by the Xiamen University Nanqiang Outstanding Talents Program. *(Corresponding author: Xiaobo Qu.)*

Chunyan Xiong is with the Institute of Electromagnetics and Acoustics School of Electronic Science and Engineering, Fujian Provincial Key Laboratory of Plasma and Magnetic Resonance, Xiamen University, Xiamen 361104, China.

Chaoxing Zhang, Mengli Lu, Xiaotong Yu, Jian Cao, Zhong Chen, and Xiaobo Qu are with the School of Electronic Science and Engineering, Fujian Provincial Key Laboratory of Plasma and Magnetic Resonance, Biomedical Intelligent Cloud Research and Development Center, Xiamen University, Xiamen 361104, China (e-mail: quxiaobo@xmu.edu.cn).

Di Guo is with the School of Computer and Information Engineering, University of Technology, Xiamen 361024, China.

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2024.3353795, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2024.3353795

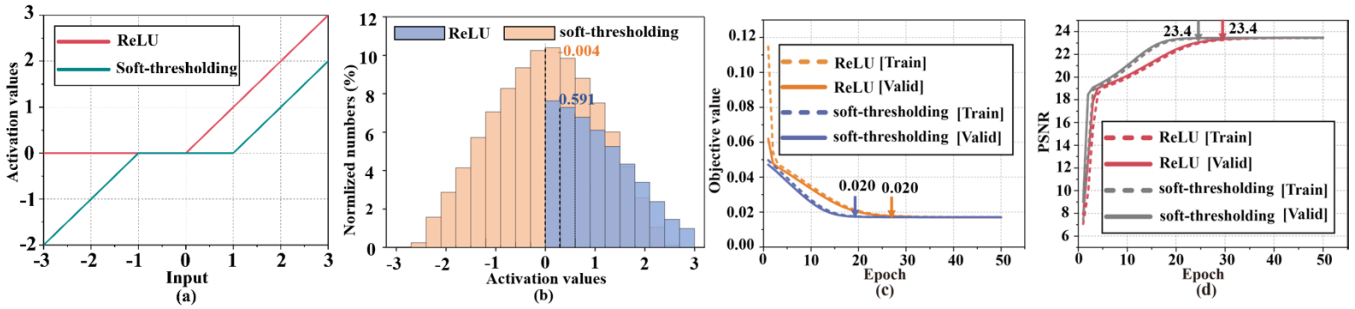


Fig. 2. Denoising performance of a two-layer convolutional NN with ReLU and ST on CIFAR-10. (a) Activate function. (b) Distribution of the activation of ST. (c) Objective value. (d) PSNR.

resources if the size of training data is large. The latter avoids this problem by reducing network complexity through local convolutions [41], [42], [43] and have been successfully applied in image processing [44], [45], [46], [47]. Up to now, CNN has been utilized as an example in convexifying networks under a common nonlinear function, ReLU [38], [39].

To make the rest of the description clear, following previous theoretical work [38], we will adopt the denoising task handled by a basic two-layer ReLU-CNN, for theoretical analysis of convexity.

Let $\widetilde{\mathbf{X}} \in \mathbb{R}^{n\times h}$ denote a noise-free 2D image, n and h represent the width and height, respectively. $\widetilde{\mathbf{X}}$ is contaminated by an additive noise $\mathbf{E}$, whose entries are drawn from a probability distribution, such as $N(0,\sigma^2)$ in the case of i.i.d Gaussian noise. Then, the noisy observation $\widetilde{\mathbf{Y}}$ is modeled as $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{X}} + \mathbf{E}$. Given a set of convolution filters, $\widetilde{\mathbf{U}}_k \in \mathbb{R}^{m\times m} (k = 1,\ldots,K)$, noise-suppressed images are obtained under each filter and then linearly combined according to [38]

$$\sum_{k=1}^{K} \left(\widetilde{\mathbf{Y}} \otimes \widetilde{\mathbf{U}}_k\right)_+ \otimes \widetilde{v}_k \tag{1}$$

where $\otimes$ represents the 2D convolution operation, $(\cdot)_+$ denotes an element-wise ReLU operation, and $\widetilde{v}_k \in \mathbb{R}$ is a 1×1 kernel used as the weight in the linear combination. Then, convolution kernels, i.e., $\widetilde{\mathbf{U}}_k$ and v_k, are obtained by minimizing the prediction loss between the noise-free image and the network output as

$$\min_{\widetilde{\mathbf{U}}_k,\widetilde{v}_k} \left\| \sum_{k=1}^{K} \left(\widetilde{\mathbf{Y}} \otimes \widetilde{\mathbf{U}}_k\right)_+ \otimes \widetilde{v}_k - \widetilde{\mathbf{X}} \right\|_F^2. \tag{2}$$

To reduce the network complexity, (2) is further improved to an object value as [38]

$$\min_{\widetilde{\mathbf{U}}_k,\widetilde{v}_k} \left\| \sum_{k=1}^{K} \left(\widetilde{\mathbf{Y}} \otimes \widetilde{\mathbf{U}}_k\right)_+ \otimes \widetilde{v}_k - \widetilde{\mathbf{X}} \right\|_F^2 + \beta \sum_{k=1}^{K} \left(\left\|\widetilde{\mathbf{U}}_k\right\|_F^2 + |\widetilde{v}_k|^2 \right) \tag{3}$$

by constraining the energy (or the power of norm) of all convolution kernels. The $\beta > 0$ is a hyperparameter to trade the prediction loss with the convolution kernel energy.

To convexify the primal network in (3), the convex duality theory was introduced to convert (3) into a dual form, enabling the reach of global minimum [38]. No gap between the primal and dual objective values has been demonstrated theoretically and experimentally [38]. This work inspired us to convexify other networks, for example, replacing ReLU with soft-thresholding (ST).

ST [48] is another nonlinear function which is expressed as

$$\tau(a_{ij})_\lambda = \begin{cases} a_{ij} + \lambda, & a_{ij} \le -\lambda \\ 0, & |a_{ij}| < -\lambda \\ a_{ij} - \lambda, & a_{ij} > -\lambda \end{cases} \tag{4}$$

where a_{ij} is an entry in a matrix (or a vector), λ is a threshold, and $|a_{ij}|$ is the absolute value of a_{ij}. If the magnitude of an entry is smaller than a threshold, this entry will be eliminated to zero. Otherwise, the magnitude will be subtracted by λ but maintains its original sign.

ST has been widely used in denoising [48], [49], [50], [51] or sub-Nyquist signal recovery through enforcing the sparsity (more zeros) [52], [53], [54], [55], [56]of a vector or low-rankness (more zeros of singular values) of a matrix [56], [57], [58], [59]. For example, the noise will be suppressed by applying the ST on the transform, e.g., wavelets, of a noisy image [60]. But different from ReLU where all negative values are zero, ST can retain useful negative features. For example, the distribution of the mean of activation of ST is closer to zero than that of ReLU, resulting in faster convergence in network training (Fig. 2). In the NN, ST is applied to remove noise that is mixed in the feature maps of a noisy image [7], [59], [61], [62], [63], [64], [65].

However, the convexity theoretical analysis of ST network is more challenging for the ST is a three-stage function [Fig. 2(a)]. Three-segment functions make it difficult to convert nonlinear operations to linear operations and keep the values constant. The values after the ST operation belong to real numbers (no fixed constraints) and constraints cannot be uniformly added to keep the values constant like ReLU. We solved this problem by adding constraints in segments in the proof process, resulting in the constraint being increased by three times compared with ReLU. Meanwhile, the convexity and strong duality of ST networks become very complicated to prove theoretically and verify experimentally.

To the best of our knowledge, the convex form of primal Two-Layer Convolutional NNs with ST (primal ST-CNN) has not been set up. This work is to design its convex structure and provide the theoretical analysis (Fig. 3). First, we derive the weak duality of the primal ST-CNN using the Lagrange

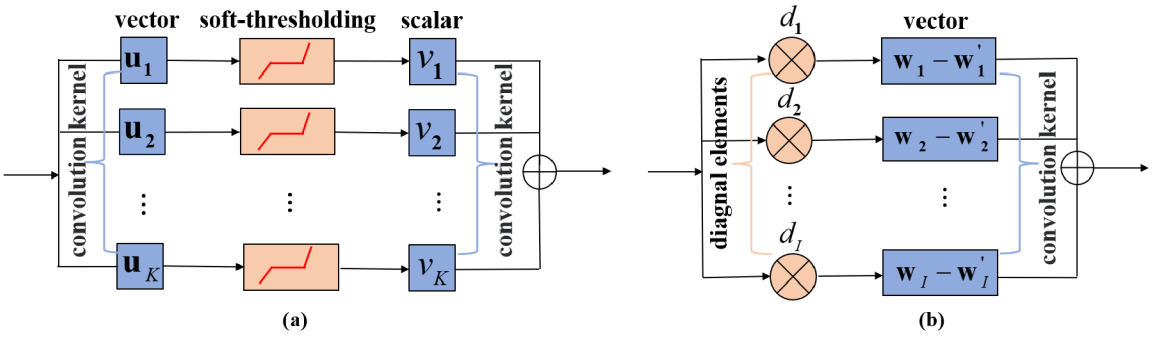


Fig. 3. NN structures. (a) Primal ST-CNN. (b) Dual ST-CNN.

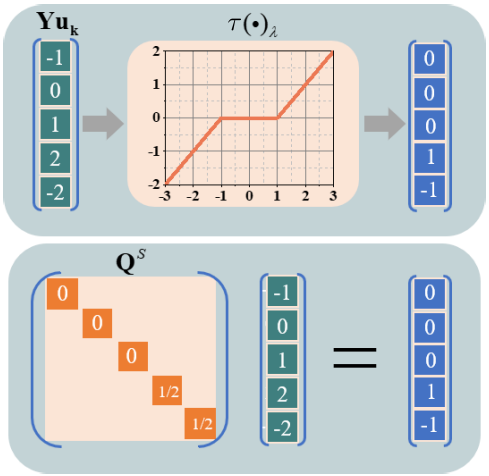


Fig. 4. Toy example: Converting nonlinear operation to linear operation. (a) ST in the primal ST-CNN. (b) Diagonal matrix in the dual ST-CNN.

dual theory. Second, the nonlinear operation is converted into a linear operation (Fig. 4), which is used to divide hyperplanes and provide exact representations in the training process. Third, we theoretically prove that the strong duality holds between the two-layer primal ST-CNN [Fig. 3(a)] and its dual form (dual ST-CNN) [Fig. 3(b)]. Fourth, experiments on both simulation and real-world datasets are conducted to support theoretical findings. Finally, the convex dual network model of a deep ST network with parallel structure is deduced theoretically.

The rest of this article is organized as follows. Section II introduces preliminaries. Section III presents the main theorem. Section IV shows experimental results and Section V makes the conclusion.

II. Preliminaries

A. Notations

Matrices and vectors are denoted by uppercase and lowercase bold letters, respectively. $\|\cdot\|_2$ and $\|\cdot\|_F$ represents Euclidean and Frobenius norms, respectively. We partition $P_s \subset \mathbb{R}^{m^2}$ into the following subsets:

$$P_s = l_1 \cup l_2 \cup l_3 \tag{5}$$

where

$$\begin{aligned} l_1 &= \{\mathbf{u}_k |\ \mathbf{y}_i^\top \mathbf{u}_k \le -\lambda\}, & H_1 &= \{i |\ \mathbf{y}_i^\top \mathbf{u}_k \le -\lambda\} \\ l_2 &= \{\mathbf{u}_k |\ -\lambda \le \mathbf{y}_i^\top \mathbf{u}_k \le \lambda\}, & H_2 &= \{i |\ -\lambda \le \mathbf{y}_i^\top \mathbf{u}_k \le \lambda\} \\ l_3 &= \{\mathbf{u}_k |\ \mathbf{y}_i^\top \mathbf{u}_k \ge \lambda\}, & H_3 &= \{i |\ \mathbf{y}_i^\top \mathbf{u}_k \ge \lambda \\ & \{\mathbf{y}_i \in \mathbb{R}^{m^2}\}_{i=1}^I, \quad I = nh, \ \mathbf{u}_k \in \mathbb{R}^{m^2}, \ \lambda \in \mathbb{R}. \end{aligned} \tag{6}$$

We denote

$$S = S_1 \cup S_2 \cup S_3 \tag{7}$$

where

$$\begin{aligned} S_1 &= \{i |\ i \in H_1\} \cup \{i |\ i \in H_2\} \\ S_2 &= \{i |\ i \in H_2\}, \\ S_3 &= \{i |\ i \in H_2\} \cup \{i |\ i \in H_3\}. \end{aligned} \tag{8}$$

$\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_I]^\top \in \mathbb{R}^{I \times m^2}$, $\mathbf{Q}^S \in \mathbb{R}^{I \times I}$ is a diagonal matrix, its diagonal elements are as follows:

$$\mathbf{Q}_{ii} = \begin{cases} \dfrac{\mathbf{y}_i^\top \mathbf{u}_k + \lambda}{\mathbf{y}_i^\top \mathbf{u}_k}, & \text{if } i \in S_1 \\ 0, & \text{if } i \in S_2 \\ \dfrac{\mathbf{y}_i^\top \mathbf{u}_k - \lambda}{\mathbf{y}_i^\top \mathbf{u}_k}, & \text{if } i \in S_3. \end{cases} \tag{9}$$

$$\mathbf{Q}^S = \mathbf{Q}^{S_1} + \mathbf{Q}^{S_2} + \mathbf{Q}^{S_3}. \tag{10}$$

We denote

$$\begin{aligned} &\mathbf{Q}^S \text{ in } \mathrm{Q}^S\mathbf{Y}\mathbf{u}_k \ge 0 \text{ as } \mathbf{Q}^{S_1} \\ &\mathbf{Q}^S \text{ in } \mathbf{Q}^S\mathbf{Y}\mathbf{u}_k = 0 \text{ as } \mathbf{Q}^{S_2} \\ &\mathbf{Q}^S \text{ in } \mathbf{Q}^S\mathbf{Y}\mathbf{u}_k \le 0 \text{ as } \mathbf{Q}^{S_2} \\ &P_S = \{\mathbf{u}_k |\ P_{S_1} \cup P_{S_2} \cup P_{S_3}\} \end{aligned} \tag{11}$$

where

$$\begin{aligned} P_{S_1} &= \{\mathbf{u}_k |\ \mathbf{Q}^{S_1}\mathbf{Y}\mathbf{u}_k \ge 0, \forall i \in S_1\} \\ P_{S_2} &= \{\mathbf{u}_k |\ \mathbf{Q}^{S_2}\mathbf{Y}\mathbf{u}_k = 0, \forall i \in S_2\} \\ P_{S_3} &= \{\mathbf{u}_k |\ \mathbf{Q}^{S_3}\mathbf{Y}\mathbf{u}_k \le 0, \forall i \in S_3\}. \end{aligned} \tag{12}$$

B. Basic Lemmas and Definitions

Lemma 1 (Slater's Condition [66]): Consider the optimization problem

$$\begin{aligned} &\min_{\mathbf{x}} \ f_0(\mathbf{x}) \\ &\text{s.t. } f_j(\mathbf{x}) < 0, \quad j = 1, \ldots, J, \quad \mathbf{A}\mathbf{x} = \mathbf{b} \end{aligned} \tag{13}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$, $f_0, \ldots, f_J$ are convex functions.

If there exists an $\mathbf{x}^* \in \mathbf{relint}D$ (where **relint** denotes the relative interior of the convex set $D := \cap_{j=0}^{J}\mathrm{dom}(f_j)$), such that

$$f_j(\mathbf{x}^*) < 0, \quad j = 0, \ldots, J, \ \mathbf{A}\mathbf{x}^* = \mathbf{b}. \tag{14}$$

Such a point is called strictly feasible since the inequality constraints hold with strict inequalities. The strong duality holds if Slater's condition holds (and the problem is convex).

Lemma 2 (Sion's Minimax Theorem [67], [68]): Let X and Y be nonvoid convex and compact subsets of two linear topological spaces, and let $f : X \times Y \to \mathbb{R}$ be a function that is upper semicontinuous and quasiconcave in the first variable and lower semicontinuous and quasiconvex in the second variable. Then,

$$\min_{y\in Y}\max_{x\in X} f(x, y) = \max_{x\in X}\min_{y\in Y} f(x, y). \tag{15}$$

Lemma 3 (Semi-infinite Programming [69]): Semi-infinite programming problems of the form

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}), \quad \text{s.t. } g(\mathbf{x}, w) \le 0, \ w \in \Omega \tag{16}$$

where Ω is a (possibly infinite) index set, $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$ denotes the extended real line, $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $g : \mathbb{R}^n \times \Omega \to \mathbb{R}$. The above optimization problem is performed in the finite-dimensional space $\mathbb{R}$ and, if the index set Ω is infinite, is subject to an infinite number of constraints, therefore, it is referred to as a semi-infinite programming problem.

Lemma 4 (An Extension of Zaslavsky's Hyperplane Arrangement Theory [70]): Consider a deep rectifier network with L layers, n_l rectified linear units at each layer l, and an input of dimension n_0. The maximal number of regions of this NN is at most

$$\sum_{(j_1,\ldots,j_L)}\prod_{l=1}^{L}\binom{n_l}{j_l} \tag{17}$$

where $J = \{(j_1, \ldots, j_L) \in \mathbb{Z}^L : 0 \le j_l \le \min\{n_0, n_1 - j_1, \ldots, n_{l-1} - j_{l-1}, n_l\}, \ \forall 1 = 1, \ldots, L\}$. This bound is tight when $L = 1$.

Definition 1 (Optimal Duality Gap [66]): The optimal value of the Lagrange dual problem is denoted as d^*, and the optimal value of the primal problem is denoted as p^*. The weak duality is defined as d^* is the best lower bound of p^* as follows:

$$d^* \le p^*. \tag{18}$$

The difference $p^* - d^*$ is called the optimal duality gap of the primal problem.

Definition 2 (Zero Duality Gap [66]): If the equality

$$d^* = p^* \tag{19}$$

holds, i.e., the optimal duality gap is zero, then we say that the strong duality holds. Strong duality means that a best bound, which can be obtained from the Lagrange dual function, is tight.

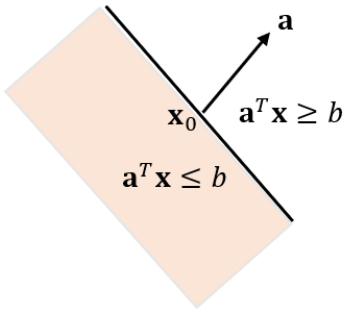


Fig. 5. Hyperplane defined by $\mathbf{a}^\top\mathbf{x} = b$ in $\mathbb{R}^2$ determines two half-spaces. The half-space determined by $\mathbf{a}^\top\mathbf{x} \ge b$ is the half-space extending in the direction $\mathbf{a}$. The half-space determined by $\mathbf{a}^\top\mathbf{x} \le b$ extends in the direction $-\mathbf{a}$. The vector $\mathbf{a}$ is the outward of this half-space.

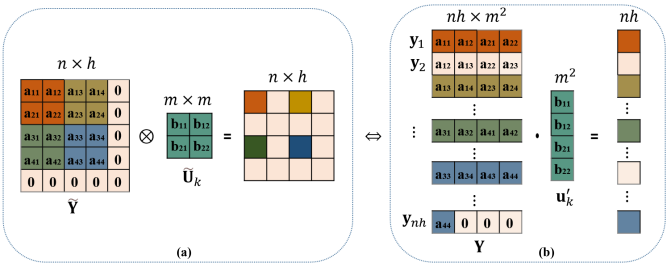


Fig. 6. Replacing convolutional operations with matrix multiplication. (a) Convolution in (20). (b) Matrix multiplication in (21).

Definition 3 (Hyperplanes and Half-spaces [66]): A hyperplane is a set of form $\{\mathbf{x}|\ \mathbf{a}^\top\mathbf{x} = b\}$ where $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{x} \in \mathbb{R}^{n\times 1}$, $\mathbf{a} \ne 0$ and $b \in \mathbb{R}$.

A hyperplane divides $\mathbb{R}^n$ into half-spaces. A half-space is a set of the form $\{\mathbf{x}|\ \mathbf{a}^\top\mathbf{x} \le b\}$ where $\mathbf{a} \ne 0$, i.e., the solution set of one (nontrivial) linear inequality. This is illustrated in Fig. 5.

III. Model and Theory

A. Proposed Model

A two-layer primal ST-CNN is expressed as follows:

$$p^* = \min_{\widetilde{\mathbf{U}}_k,\widetilde{v}_k}\left\|\sum_{k=1}^{K}\tau\left(\widetilde{\mathbf{Y}}\otimes\widetilde{\mathbf{U}}_k\right)_\lambda \otimes \widetilde{v}_k - \widetilde{\mathbf{X}}\right\|_F^2 + \beta\sum_{k=1}^{K}\left(\left\|\widetilde{\mathbf{U}}_k\right\|_F^2 + |\widetilde{v}_k|^2\right) \tag{20}$$

where the main difference between (20) and (3) is an element-wise ST operator $\tau(a_{ij})_\lambda = (|a_{ij}| - \lambda)_+\mathrm{sign}(a_{ij})$, $\otimes$ represents the 2D convolution operation, $\widetilde{\mathbf{Y}} \in \mathbb{R}^{n\times h}$ is the input, $\widetilde{\mathbf{U}}_k \in \mathbb{R}^{m\times m}$, $\widetilde{v}_k \in \mathbb{R}$, $\beta > 0$.

Replacing convolutional operations with matrix multiplication (Fig. 6), (20) can be converted into the following form:

$$p^* = \min_{\mathbf{u}'_k,v'_k}\left\|\sum_{k=1}^{K}\tau\left(\mathbf{Y}\mathbf{u}'_k\right)_\lambda v'_k - \mathbf{x}\right\|_2^2 + \beta\sum_{k=1}^{K}\left(\left\|\mathbf{u}'_k\right\|_2^2 + |v'_k|^2\right) \tag{21}$$

where $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_I]^\top \in \mathbb{R}^{I\times m^2}$ is the input, $I = nh$, $\{\mathbf{y}_i \in \mathbb{R}^{m^2}\}_{i=1}^{I}$, $\mathbf{x} \in \mathbb{R}^I$ is the label, $\mathbf{u}'_k \in \mathbb{R}^{m^2}$, $v'_k \in \mathbb{R}$.

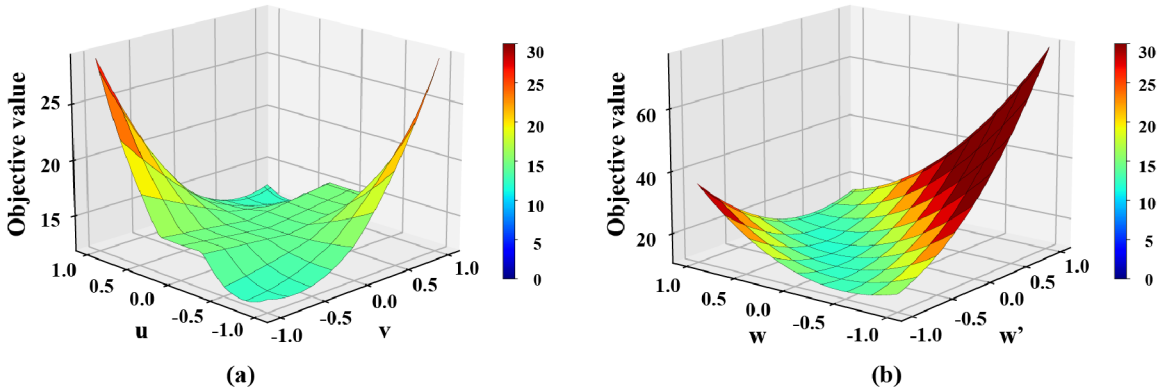


Fig. 7. Objective value of a two-layer primal ST-CNN and dual ST-CNN trained with ADAM on a 1-D dataset. Assuming $\mathbf{x} = [-1, 2, 0, 1, 2]^\top$ and $\mathbf{y} = [2, 1, 2, 1, 2]^\top$, which are the input and output, respectively. (a) Nonconvex primal ST-CNN. (b) Convex dual ST-CNN.

Next, we introduce the main theory (Theorem 1) that converts a two-layer primal ST-CNN [Fig. 7(a)] into a convex dual ST-CNN [Fig. 7(b)].

B. Theoretical Analysis

Theorem 1 (Main Theory): There exists $k^* \leq I$ such that if the number of convolution filters $k \geq (k^* + 1)$, a two-layer ST-CNN (21) has a strong duality satisfy form. This form is given through finite-dimensional convex programming as

$$d_3^* = \min_{\mathbf{w}_i \in p_w, \mathbf{w}_i' \in p_{w'}} \left\| \sum_{i=1}^{I} \mathbf{Q}^S \mathbf{Y}(\mathbf{w}_i' - \mathbf{w}_i) - \mathbf{x} \right\|_2^2 + 2\beta \sum_{i=1}^{I} \left(\|\mathbf{w}_i'\|_2 + \|\mathbf{w}_i\|_2 \right) \quad (22)$$

where $\mathbf{Q}^S$ is a diagonal matrix, and its diagonal elements for $\mathbf{Q}_{ii}$ take the following values:

$$\mathbf{Q}_{ii} = \begin{cases} \dfrac{\mathbf{y}_i^\top \mathbf{u}_k + \lambda}{\mathbf{y}_i^\top \mathbf{u}_k}, & \text{if } i \in S_1 \\ 0, & \text{if } i \in S_2 \\ \dfrac{\mathbf{y}_i^\top \mathbf{u}_k - \lambda}{\mathbf{y}_i^\top \mathbf{u}_k}, & \text{if } i \in S_3. \end{cases} \quad (23)$$

$\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_I]^\top \in \mathbb{R}^{I \times m^2}$, $\mathbf{w}_i$ and $\mathbf{w}_i'$ are both dual variables, and they correspond to $\mathbf{u}_k'$ and v_k' in (21) which are learnable parameters. $\mathbf{x} \in \mathbb{R}^I$ is the label

$$p_w = \left\{ \mathbf{w}_i | \mathbf{Q}^{S_1} \mathbf{Y} \mathbf{w}_i \geq 0, \mathbf{Q}^{S_2} \mathbf{Y} \mathbf{w}_i = 0, \mathbf{Q}^{S_3} \mathbf{Y} \mathbf{w}_i \leq 0 \right\}$$
$$p_{w'} = \left\{ \mathbf{w}_i' | \mathbf{Q}^{S_1} \mathbf{Y} \mathbf{w}_i' \geq 0, \mathbf{Q}^{S_2} \mathbf{Y} \mathbf{w}_i' = 0, \mathbf{Q}^{S_3} \mathbf{Y} \mathbf{w}_i' \leq 0 \right\}. \quad (24)$$
$$\mathbf{Q}^S = \mathbf{Q}^{S_1} + \mathbf{Q}^{S_2} + \mathbf{Q}^{S_3}. \quad (25)$$

Remark: The constraints on $\mathbf{w}$ and $\mathbf{w}'$ in p_w and p_w' arise from the segmentation property of the soft thresholding. We first randomly generate the vector $\bar{\mathbf{w}}$ to do convolution with the input $\mathbf{Y}$ and generate the corresponding $\mathbf{Q}^S$ based on the value of $\mathbf{Y}\bar{\mathbf{w}}$. Then, we input $\mathbf{Y}$ and $\mathbf{Q}^S$ into our dual ST-CNN, and (22) is used in our objective function (objective loss). Because it is an objective function with constraints p_w and p_w', hence, we use hinge loss (adding constraints to the objective function) as the loss function in experiments. There exist $\mathbf{w}_i$, $\mathbf{w}_i'$ such that $\bar{\mathbf{w}} = \mathbf{w}_i' - \mathbf{w}_i$.

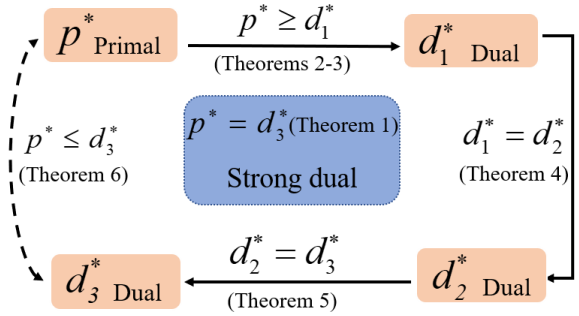


Fig. 8. Main derivation process.

As an extension, dual theory analysis of three-layer ST subnetworks has been proved (see Supplementary Material E).

Before proving the main theory (Theorem 1), we present the following main derivation framework (Fig. 8). The strong duality relationship between the primal network and the dual network is established, meaning that $p^* \to d_1$, $d_1 \to d_2$, $d_2 \to d_3$ can be achieved at the same time.

1) *Theorem 2:* Scaling $\|\mathbf{u}_k'\|_2^2 + |v_k'|^2$ in the primal ST-CNN (21).
2) *Theorem 3:* Eliminating variables to obtain an equivalent convex optimization model under the principle of Lagrangian dual theory.
3) *Theorem 4:* Convert nonlinear operations to linear operations using a diagonal matrix.
4) *Theorem 5:* Exact representation of a two-layer ST-CNN.
5) *Theorem 6:* Prove zero dual gaps (strong duality).

Theorem 2: To scaling $\mathbf{u}_k'$, v_k', let $\mathbf{u}_k = \varepsilon \mathbf{u}_k'$, $v_k = \frac{1}{\varepsilon} v'_k$

$$p^* = \min_{\mathbf{u}_k', v_k'} \left\| \sum_{k=1}^{K} \tau\left(\mathbf{Y}\mathbf{u}_k'\right)_\lambda v_k' - \mathbf{x} \right\|_2^2 + \beta \sum_{k=1}^{K} \left(\|\mathbf{u}_k'\|_2^2 + |v_k'|^2 \right) \quad (26)$$

the primal ST-CNN can be translated as

$$p^* = \min_{\|\mathbf{u}_k\|_2 \leq 1} \min_{v_k \in \mathbb{R}} \left\| \sum_{k=1}^{K} \tau(\mathbf{Y}\mathbf{u}_k)_\lambda v_k - \mathbf{x} \right\|_2^2 + 2\beta \sum_{k=1}^{K} (|v_k|) \quad (27)$$

where ε is introduced so that the scaling has no effect on the network output, the proof of Theorem 2 is provided in Supplementary Material A.

Then, according to (27), we can obtain an equivalent convex optimization model by using the Lagrangian dual theory.

Theorem 3:

$$p^* = \min_{\|\mathbf{u}_k\|_2\le 1} \min_{v_k\in\mathbb{R}} \left\| \sum_{k=1}^{K} \tau(\mathbf{Y}\mathbf{u}_k)_\lambda v_k - \mathbf{x} \right\|_2^2 + 2\beta \sum_{k=1}^{K} |v_k|$$

is equivalent to

$$d_1^* = \max_{\|\mathbf{u}_k\|_2\le 1, \mathbf{z}:|\mathbf{z}^\top \tau(\mathbf{Y}\mathbf{u}_k)_\lambda|\le 2\beta} -\frac{1}{4}\|\mathbf{z}-2\mathbf{x}\|_2^2 + \|\mathbf{x}\|_2^2. \quad (28)$$

Proof: By reparameterizing the problem, let

$$\mathbf{r} = \sum_{k=1}^{K} \tau(\mathbf{Y}\mathbf{u}_k)_\lambda v_k - \mathbf{x} \quad (29)$$

where $\mathbf{r}\in\mathbb{R}^I$, hence, we have

$$d_1^* = \min_{\|\mathbf{u}_k\|_2\le 1} \min_{v_k,\mathbf{r}} \|\mathbf{r}\|_2^2 + 2\beta \sum_{k=1}^{K} |v_k|$$
$$\text{s.t. } \mathbf{r} = \sum_{k=1}^{K} \tau(\mathbf{Y}\mathbf{u}_k)_\lambda v_k - \mathbf{x}. \quad (30)$$

Introducing the Lagrangian variable $\mathbf{z}$, and $\mathbf{z}\in\mathbb{R}^I$, $\mathbf{z}^\top \in \mathbb{R}^{1\times I}$, and obtaining the Lagrangian dual form of the primal ST-CNN as follows:

$$d_1^* = \min_{\|\mathbf{u}_k\|_2\le 1} \min_{v_k,\mathbf{r}} \max_{\mathbf{z}} \|\mathbf{r}\|_2^2 + 2\beta \sum_{k=1}^{K} |v_k| + \mathbf{z}^\top \mathbf{r} + \mathbf{z}^\top \mathbf{x} - \mathbf{z}^\top \sum_{k=1}^{K} \tau(\mathbf{Y}\mathbf{u}_k)_\lambda v_k. \quad (31)$$

Using Sion's minimax theorem [67], [68] to change the order of maximum and minimum

$$d_1^* = \min_{\|\mathbf{u}_k\|_2\le 1} \max_{\mathbf{z}} \min_{v_k,\mathbf{r}} \|\mathbf{r}\|_2^2 + 2\beta \sum_{k=1}^{K} |v_k| + \mathbf{z}^\top \mathbf{r} + \mathbf{z}^\top \mathbf{x} - \mathbf{z}^\top \sum_{k=1}^{K} \tau(\mathbf{Y}\mathbf{u}_k)_\lambda v_k. \quad (32)$$

Minimizing the objective function (32) with $\mathbf{r}$ as a variable

$$\|\mathbf{r}\|_2^2 + \mathbf{z}^\top \mathbf{r} = \left\| \mathbf{r} + \frac{1}{2}\mathbf{z} \right\|_2^2 - \frac{1}{4}\|\mathbf{z}\|_2^2. \quad (33)$$

When $\mathbf{r} = -\frac{1}{2}\mathbf{z}$, (32) takes the optimal value. Hence, (32) can be translated to

$$d_1^* = \min_{\|\mathbf{u}_k\|_2\le 1} \max_{\mathbf{z}} \min_{v_k} -\frac{1}{4}\|\mathbf{z}\|_2^2 + 2\beta \sum_{k=1}^{K} |v_k| + \mathbf{z}^\top \mathbf{x} - \mathbf{z}^\top \sum_{k=1}^{K} \tau(\mathbf{Y}\mathbf{u}_k)_\lambda v_k. \quad (34)$$

Let

$$f = \min_{v_k} 2\beta \sum_{k=1}^{K} |v_k| - \mathbf{z}^\top \sum_{k=1}^{K} \tau(\mathbf{Y}\mathbf{u}_k)_\lambda v_k \quad (35)$$

eliminating the variable v_k in the primal ST-CNN, hence

$$\max_{\mathbf{z}:\|\mathbf{u}_k\|_2\le 1} |\mathbf{z}^\top \sum_{k=1}^{K} \tau(\mathbf{Y}\mathbf{u}_k)_\lambda| \le 2\beta. \quad (36)$$

Equation (34) is equivalent to the following optimization problem:

$$d_1^* = \max_{\|\mathbf{u}_k\|_2\le 1, \mathbf{z}:|\mathbf{z}^\top \sum_{k=1}^{K} \tau(\mathbf{Y}\mathbf{u}_k)_\lambda|\le 2\beta} -\frac{1}{4}\|\mathbf{z}-2\mathbf{x}\|_2^2 + \|\mathbf{x}\|_2^2. \quad (37)$$

□

Next, to divide hyperplanes and provide an exact representation, we convert the nonlinear operation $\tau(\cdot)_\lambda$ into the linear operator using the diagonal matrix $\mathbf{Q}^S$.

Theorem 4:

$$d_1^* = \max_{\|\mathbf{u}_k\|_2\le 1, \mathbf{z}:|\mathbf{z}^\top \sum_{k=1}^{K} \tau(\mathbf{Y}\mathbf{u}_k)_\lambda|\le 2\beta} -\frac{1}{4}\|\mathbf{z}-2\mathbf{x}\|_2^2 + \|\mathbf{x}\|_2^2$$

can be represented as a standard finite-dimensional program

$$d_2^* = \max_{\mathbf{z}} -\frac{1}{4}\|\mathbf{z}-2\mathbf{x}\|_2^2 + \|\mathbf{x}\|_2^2 \quad (38)$$
$$\text{s.t. } P_S = \{\mathbf{u}_k |\ P_{S_1} \cup P_{S_2} \cup P_{S_3}\} \quad (39)$$

where

$$P_{S_1} = \{\mathbf{u}_k |\ \mathbf{Q}^{S_1}\mathbf{Y}\mathbf{u}_k \ge 0, \forall i \in S_1\}$$
$$P_{S_2} = \{\mathbf{u}_k |\ \mathbf{Q}^{S_2}\mathbf{Y}\mathbf{u}_k = 0, \forall i \in S_2\}$$
$$P_{S_3} = \{\mathbf{u}_k |\ \mathbf{Q}^{S_3}\mathbf{Y}\mathbf{u}_k \le 0, \forall i \in S_3\}. \quad (40)$$

Proof: First, we analyze the one-sided dual constraint in (36) as follows:

$$\max_{\mathbf{z}:\|\mathbf{u}_k\|_2\le 1} \mathbf{z}^\top \tau(\mathbf{Y}\mathbf{u}_k)_\lambda \le 2\beta. \quad (41)$$

To divide hyperplanes, we divide $\mathbb{R}^{m^2}$ into three subsets to obtain (5) and (6). Let $i \in H_1 \cup H_2 \cup H_3$, $|H_1|+|H_2|+|H_3| = nh = I$, $\mathcal{H}_X$ be the set of all hyperplane arrangement patterns for the matrix $\mathbf{Y}$, defined as the following set [71], [72]

$$\mathcal{H}_X = \{\text{sign}(\mathbf{Y}\mathbf{u}_k + \lambda) \cup \text{sign}(\mathbf{Y}\mathbf{u}_k - \lambda) | \mathbf{u}_k \in \mathbb{R}^{m^2}\}. \quad (42)$$

Next, we take out the positions of the elements corresponding to different symbols and assign them according to

$$S_1 = \{i |\ i \in H_1\} \cup \{i |\ i \in H_2\}$$
$$S_2 = \{i |\ i \in H_2\}$$
$$S_3 = \{i |\ i \in H_2\} \cup \{i |\ i \in H_3\}$$
$$S = S_1 \cup S_2 \cup S_3. \quad (43)$$

To assign a corresponding value to the position of each i in the above three sets such that the same transformation as the soft threshold function is achieved, the diagonal matrix $\mathbf{Q}^S$ is constructed, and its diagonal elements for $\mathbf{Q}_{ii}$ as (9).

Using the diagonal matrix $\mathbf{Q}^S$, the constraints in (36) are equivalent to the following form:

$$\max_{\|\mathbf{u}_k\|_2 \le 1,\ P_S} |\mathbf{z}^\top \mathbf{Q}^S (\mathbf{Y}\mathbf{u}_k)| \le 2\beta \quad (44)$$

where $\mathbf{Q}^S = \mathbf{Q}^{S_1} + \mathbf{Q}^{S_2} + \mathbf{Q}^{S_3}$.

Hence, (37) can be finitely parameterized as

$$d_2^* = \max_{\mathbf{z}} -\frac{1}{4}\|\mathbf{z}-2\mathbf{x}\|_2^2 + \|\mathbf{x}\|_2^2 \quad \text{s.t.} \quad \max_{\|\mathbf{u}_k\|_2 \le 1,\, P_S} |\mathbf{z}^\top \mathbf{Q}^S \mathbf{Y}\mathbf{u}_k| \le 2\beta. \tag{45}$$

□

Now, we introduce an exact representation of a two-layer ST-CNN.

Theorem 5:

$$d_2^* = \max_{\mathbf{z}} -\frac{1}{4}\|\mathbf{z}-2\mathbf{x}\|_2^2 + \|\mathbf{x}\|_2^2 \quad \text{s.t.} \quad \max_{\|\mathbf{u}_k\|_2 \le 1,\, P_S} |\mathbf{z}^\top \mathbf{Q}^S \mathbf{Y}\mathbf{u}_k| \le 2\beta$$

is equivalent to

$$d_3^* = \min_{\mathbf{w}_i \in p_w, \mathbf{w}_i' \in p_{w'}} \left\| \sum_{i=1}^{I} \mathbf{Q}^S \mathbf{Y}(\mathbf{w}_i' - \mathbf{w}_i) - \mathbf{x} \right\|_2^2 + 2\beta \sum_{i=1}^{I} (\|\mathbf{w}_i\|_2 + \|\mathbf{w}_i'\|_2) \tag{46}$$

where

$$p_w = \left\{ \mathbf{w}_i | \mathbf{Q}^{S_1} \mathbf{Y}\mathbf{w}_i \ge 0, \mathbf{Q}^{S_2} \mathbf{Y}\mathbf{w}_i = 0, \mathbf{Q}^{S_3} \mathbf{Y}\mathbf{w}_i \le 0 \right\}$$
$$p_{w'} = \left\{ \mathbf{w}_i' | \mathbf{Q}^{S_1} \mathbf{Y}\ \mathbf{w}_i' \ge 0 \mathbf{Q}^{S_2} \mathbf{Y}\mathbf{w}_i' = 0, \mathbf{Q}^{S_3} \mathbf{Y}\mathbf{w}_i' \le 0 \right\}.$$

The proof of Theorem 5 is provided in Supplementary Material B. According to this theorem, we can prove that the strong duality holds, i.e., the primal ST-CNN and the dual ST-CNN achieve global optimality. They are theoretically equivalent and will obtain Theorem 6.

Theorem 6: Suppose the optimal value of the primal ST-CNN is p^* and the optimal value of the dual ST-CNN is d_3^*, the strong duality holds if $p^* = d_3^*$.

Proof: The optimal solution to the dual ST-CNN is the same as the optimal solution to the primal ST-CNN model constructed $\{\mathbf{u}_k^*, v_k^*\}_{k=1}^K$ as follows:

$$(\mathbf{u}_k^*, v_k^*) = \left(\frac{\mathbf{w}_i^*}{\sqrt{\|\mathbf{w}_i^*\|}}, \sqrt{\|\mathbf{w}_i^*\|} \right), \quad \text{if } \mathbf{w}_i^* \neq 0$$
$$(\mathbf{u}_k^*, v_k^*) = \left(\frac{\mathbf{w}_i'^*}{\sqrt{\|\mathbf{w}_i'^*\|}}, \sqrt{\|\mathbf{w}_i'^*\|} \right), \quad \text{if } \mathbf{w}_i'^* \neq 0 \tag{47}$$

where $\{\mathbf{w}_i^*, \mathbf{w}_i'^*\}_{i=1}^I$ are the optimal solution of (46)

$$\begin{aligned} p^* &= \min_{\mathbf{u}_k' \in \mathbb{R}^m, v_k' \in \mathbb{R}} \left\| \sum_{k=1}^{K} \tau(\mathbf{Y}\mathbf{u}_k')_\lambda v_k' - \mathbf{x} \right\|_2^2 + \beta \sum_{k=1}^{K} \left(\|\mathbf{u}_k'\|_2^2 + |v_k'|^2 \right) \\ &\le \left\| \sum_{k=1}^{K} \tau(\mathbf{Y}\mathbf{u}_k^*)_\lambda v_k^* - \mathbf{x} \right\|_2^2 + \beta \sum_{k=1}^{I} (\|\mathbf{u}_k^*\|_2^2 + |v_k^*|^2) \\ &= \left\| \sum_{i=1}^{I} \mathbf{Q}^S \mathbf{Y}(\mathbf{w}_i' - \mathbf{w}_i) - \mathbf{x} \right\|_2^2 \\ &\quad + \beta \sum_{i=1, \mathbf{w}_i^* \neq 0}^{I} \left(\left\| \frac{\mathbf{w}_i^*}{\sqrt{\|\mathbf{w}_i^*\|_2}} \right\|_2^2 + \left\| \sqrt{\|\mathbf{w}_i^*\|_2} \right\|_2^2 \right) \\ &\quad + \beta \sum_{i=1, \mathbf{w}_i^* \neq 0}^{I} \left(\left\| \frac{\mathbf{w}_i'^*}{\sqrt{\|\mathbf{w}_i'^*\|_2}} \right\|_2^2 + \left\| \sqrt{\|\mathbf{w}_i'^*\|_2} \right\|_2^2 \right) \\ &= \left\| \sum_{i=1}^{I} \mathbf{Q}^S \mathbf{Y}(\mathbf{w}_i'^* - \mathbf{w}_i^*) - \mathbf{x} \right\|_2^2 \\ &\quad + 2\beta \sum_{i=1}^{I} \left(\|\mathbf{w}_i^*\|_2 + \|\mathbf{w}_i'^*\|_2 \right) = d_3^*. \end{aligned} \tag{48}$$

Combining $p^* \le d_3^*$, $p^* \ge d_1^*$ (Theorems 2–3) and $d_1^* = d_2^* = d_3^*$ (Theorems 4–5), $p^* = d_3^*$ is proved.

Basing on the Lemma 3 [69], we know that $k+1$ of the total I filters $(\mathbf{w}_i, \mathbf{w}_i')$ are nonzero at optimum, where $k \le I$ [38], [39].

Finally, by combining Theorems 2–6, the main theory can be proved. Thus, the hyperplane arrangements can be constructed in polynomial time (See proof in Supplementary Material C).

It is useful to recognize that two-layer ST networks with K hidden neurons can be globally optimized via the convex program [see (22)]. The convex program has $6I^2$ constraints and $6Im^2$ variables, which can be solved in polynomial time with respect to I. The computational complexity is at most $O(m^{12}(I/m^2)^{3m^2})$ using standard interior-point solvers.

The global optimization of NNs is NP-Hard [73]. Despite the theoretical difficulty, highly accurate models are trained in the practice using stochastic gradient methods [74]. Unfortunately, stochastic gradient methods cannot guarantee convergence to an optimum of the nonconvex training loss [75] and existing methods rarely certify convergence to a stationary point of any type [76]. Stochastic gradient methods are also sensitive to hyperparameters; they converge slowly to different stationary points [77] or even diverge depending on the choice of step size. Parameters like the random seed complicate replications and can produce model churn, where networks learn using the same procedure give different predictions for the same inputs [78], [79].

Therefore, some optimizers were designed to find the optimal solution during the training process. For example, early on, the SGD optimizer [80], the SGD-based adaptive gradient optimizer (ASGD) [81], stochastic gradient descent with momentum (SGDM) [82], Adaptive Gradient (AdaGrad) [83], Root Mean Square Propagation (RMSprop) [84], and the adoption of moment estimation (ADAM) [23] optimizer, may lead to different training results under the same nonconvex optimization objective [85], which are also observed in our experimental results (see Section IV).

IV. Experimental Results

Experiments will show three observations: 1) the performance of the primal ST-CNN depends on the chosen optimizer; 2) the performance of the primal ST-CNN relies

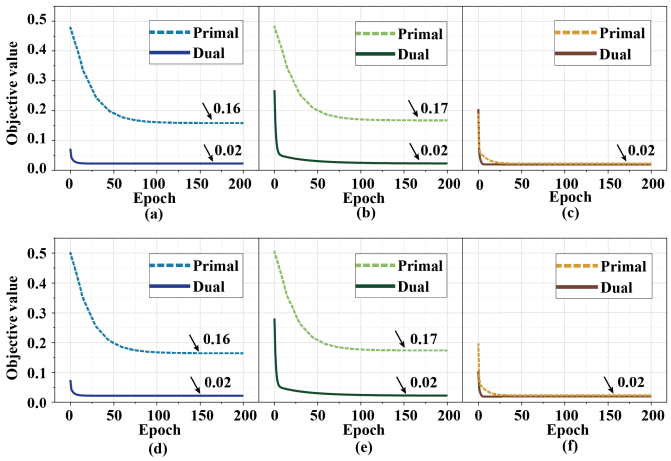


Fig. 9. Training and testing with the different optimizers. (a)–(c) are training results under the optimizer of SGD, ASGD, and Adam, respectively. (d)–(f) are test results under the optimizer of SGD, ASGD, and Adam, respectively.

on initialization; and 3) the zero dual gap holds between the primal and dual ST-CNN.

All experiments were implemented on a server equipped with dual Intel Xeon Silver 4210 CPUs, 128 GB RAM, the Nvidia Tesla T4 GPU (16 GB memory), and PyTorch deep learning library [86]. The test dataset includes simulated data, the MNIST handwritten digits, and CIFAR-10 commonly used in deep learning research [38], [87], [88].

Experiment A uses the MNIST handwritten digits dataset with a size of 28 × 28. We randomly select 600 out of 60 000 as the training set, and 10 000 in the test set remain the same. Then, they are added with i.i.d Gaussian noise from the distribution $N(0, \sigma^2)$ as the training and test dataset of the primal and the dual ST-CNN.

For network training, 600 noisy images and their noise-free ones are used as the input and label. For the network test, 10 000 noisy images and their noise-free ones are used as the input and label. The number of training and test datasets is consistent with that used in the ReLU-based dual theory experiments [38]. When verifying the strong duality, we use the training set 600 (see Supplementary Material E) and increase the training set to 6000.

Experiment B and C uses the CIFAR-10 dataset with a size of 32 × 32. We randomly select 3000 out of 50 000 as the training set, and 400 out of 10 000 as the test set. In the denoising experiment, they are added with i.i.d Gaussian noise from the normal distribution $N(0, \sigma^2)$ where 0 is the mean and σ is the standard deviation. For network training, 3000 noisy images and their noise-free ones are used as the input and label. For the network test, 400 noisy images and their noise-free ones are used as the input and label.

A. Experiment on Simulation Dataset MNIST

1) Primal ST-CNN Relies on Optimizer: Here, we choose noise $\sigma = 0.25$, and the primal ST-CNN and the dual ST-CNN are trained by using SGD, ASGD [81], and ADAM [23] as the optimizers, respectively. The training and testing results are shown in Fig. 9. The optimal solution of the primal ST-CNN is dependent on the selection of optimizers.

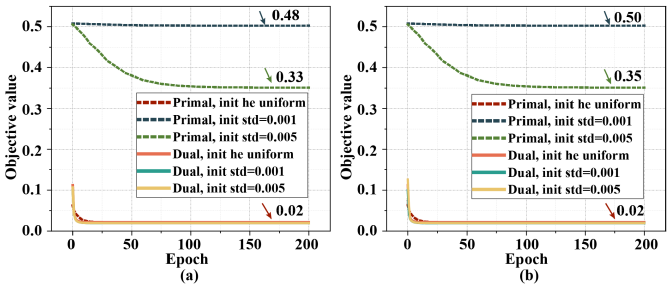


Fig. 10. Example of training and testing with the different initialization. The primal ST-CNN and the dual ST-CNN are trained with Kaiming uniform initial [16], mean 0, standard deviation 0.001, and standard deviation 0.005 initialized with normal distribution, respectively. (a) Training results. (b) Testing results.

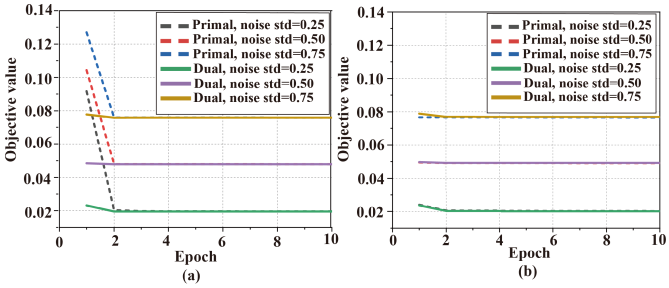


Fig. 11. Verify that zero dual gap (the strong duality) holds, i.e., the objective function values are very close when both the primal ST-CNN and the dual network objective value achieves global optimality. (a) and (b) is the objective value under the training and test stage (networks are trained), respectively.

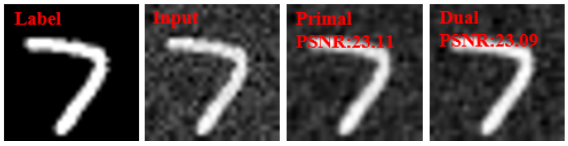


Fig. 12. Representative denoised images were obtained by the primal and dual networks. Note: Noise following a Gaussian distribution with a mean 0 and standard deviation of 0.25 is added to the MNIST image.

2) Primal ST-CNN Relies on Initialization: We choose different ways of parameter initialization including Kaiming He uniform distribution initialized as well as a normal distribution with zero mean and standard deviation of 0.001 and 0.005, respectively. The experimental results are shown in Fig. 10. The objective value of the primal ST-CNN and the dual ST-CNN coincide when two types of networks are initialized with Kaiming He uniform distribution for training. However, when we initialize the network parameters using normal distributions with mean 0 and standard deviations of 0.001 and 0.005, the objective value of the dual ST-CNN will be better than the primal ST-CNN.

This observation implies that the primal ST-CNN is dependent on the selection of the initial values.

3) Verify Zero Dual Gap (Strong Duality): Zero dual gap [66] means that, when both the primal and the dual ST-CNN reach the global optimum, the objective values of the two are equal. Therefore, according to the above two experimental results, in order to make the primal network achieve the global optimum, we choose ADAM [23] as the

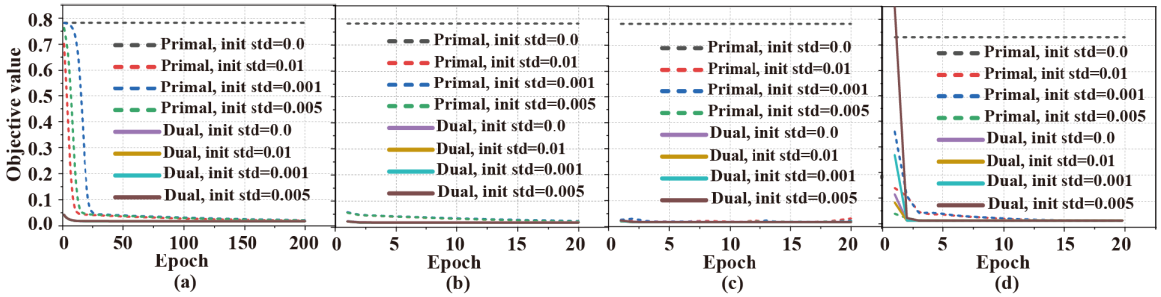


Fig. 13. Objective value of the primal and dual network under different optimizers and different initialization modes. (a)–(d) are results under the optimizer of SGDM, AdaGrad, RMSProp, and Adam, respectively.

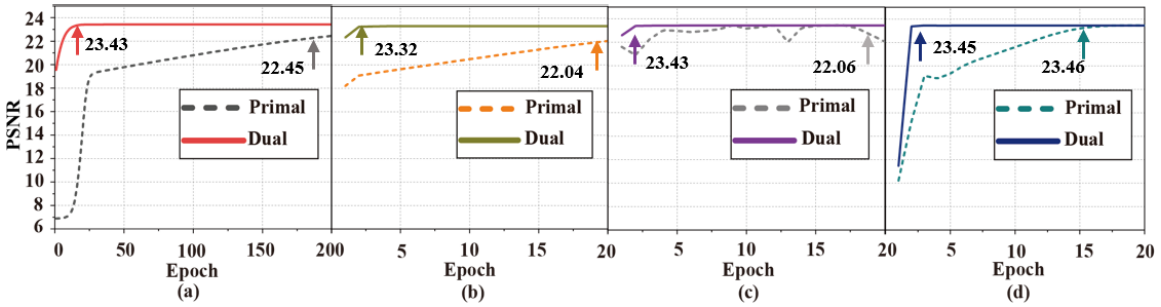


Fig. 14. PSNR performance of the primal and dual networks under different optimizers. (a)–(d) are results under the optimizer of SGDM, AdaGrad, RMSProp, and Adam, respectively. Note: The PSNR is averaged on 400 noisy images after the networks have been trained. All network parameters are initialized following a normal distribution $N(0, 0.001^2)$.

TABLE I
PSNR/RLNE OF DENOISING MNIST DATASET BY THE TWO NETWORKS

Noise standard deviation	Pri_train	Pri_test	Dual_train	Dual_test
0.25	23.09 / 0.197	23.09 / 0.205	23.11 / 0.197	23.11 / 0.205
0.50	19.20 / 0.309	19.12 / 0.324	19.21 / 0.308	19.12 / 0.324
0.75	17.20 / 0.388	17.16 / 0.406	17.21 / 0.388	17.15 / 0.407

optimizer for the primal network, and the network parameters are initialized by Kaiming He uniform distribution.

Under various noise levels, $\sigma \in \{0.25, 0.50, 0.75\}$. Both approaches achieve close objective values under all noise levels (Fig. 11).

From Fig. 12, Table I, the denoising results of the primal network and the dual network is very close. Since the experiment is based on database training, the difference of PSNR after denoising is less than 1.0, and the difference of RLNE is less than 0.1 both within the acceptable range. Therefore, the strong duality is valid.

B. Experiment on Simulation Dataset CIFAR-10

In experiment B, we first verify primal ST-CNN relies on initialization. Then, primal ST-CNN relies on the type of optimizer that is verified. Finally, we verify strong duality.

1) Primal ST-CNN Relies on Initialization: The network parameters are initialized following four normal distributions with mean 0, and standard deviations of 0.0, 0.01, 0.001, and 0.005. Under four optimizers, we test the effect of different initializations on the objective value.

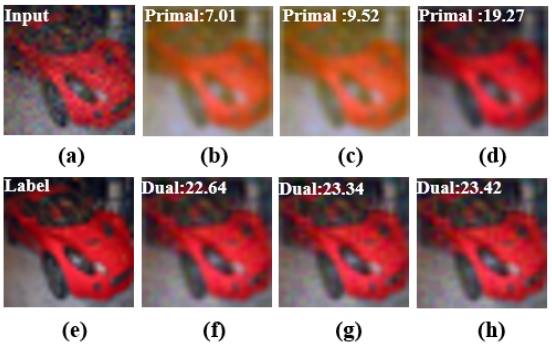


Fig. 15. Representative denoised images in different epochs. (a) and (e) is the input (noisy image) and label (noise-free image), (b)–(d) [or (f)–(h)] are results obtained by the primal (or dual) network when epoch is 8, 16, and 28, respectively. Note: SGDM optimizer and $N(0, 0.001^2)$ initializations are adopted.

Fig. 13 shows that the primal network cannot achieve global optimization when the initial standard deviation is 0.0 under the four optimizers. On the contrary, the dual network can reach the optimal value of 0.018 for all cases.

2) Primal ST-CNN Relies on Optimizer: In this part, all network parameters are initialized following a normal distribution $N(0, 0.001^2)$.

Under four optimizers, we test the denoising performance of the primal and dual networks. The experimental dataset is consistent with the above experiment. Under the optimizers of SGDM [Fig. 14(a)] and AdaGrad [Fig. 14(b)], although objective values of both primal and dual networks converge, the denoising performance peak signal-to-noise ratio (PSNR) does not reach the highest value. Representative denoised

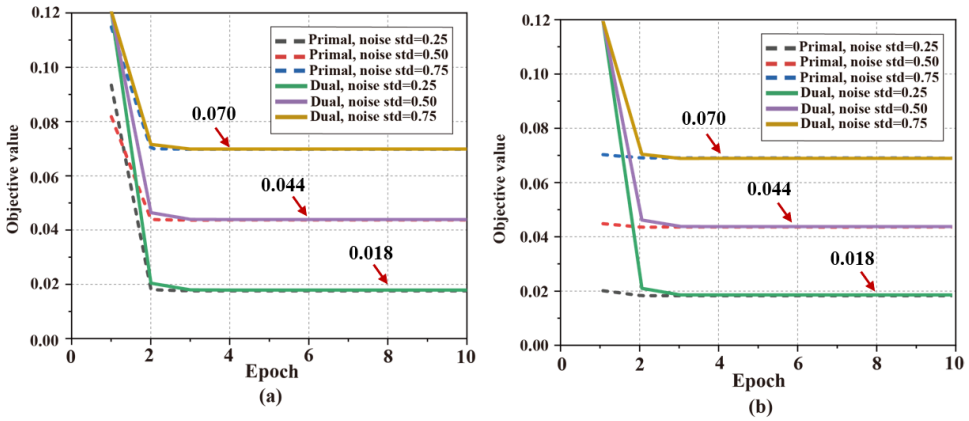


Fig. 16. Verification of the zero dual gap (the strong duality), i.e., the objective values are very close when both the primal and dual networks achieve the global optimality. (a) and (b) is the objective value under the training and test stage (networks are trained), respectively. Note: Noise is added following a Gaussian distribution with the mean 0 and the standard deviations 0.25, 0.50, and 0.75, respectively.

TABLE II

PSNR OF DENOISING CIFAR-10 DATASET BY THE TWO NETWORKS

Noise standard deviation	Pri_train	Pri_test	Dual_train	Dual_test
0.25	23.44	23.46	23.17	23.20
0.50	19.45	19.50	19.36	19.40
0.75	17.41	17.45	17.36	17.41

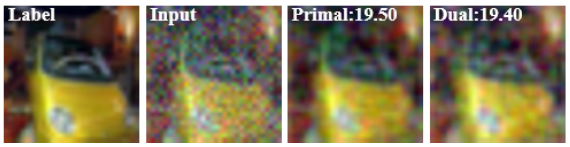


Fig. 17. Representative denoised images were obtained by the primal and dual networks. Note: Noise following a Gaussian distribution with a mean 0 and standard deviation of 0.50 is added to the CIFAR-10 image.

images are shown in Fig. 15. The intermediate and final images differ greatly from the noise-free ones when the optimizer of SGDM is applied. Under the optimizer of RMSProp [Fig. 14(c)], the PSNRs can reach the best performance but are unstable. Under the optimizer of Adam [Fig. 14(d)], the primal network converges to the best PSNR which is the same as that obtained by the dual network. Under all optimizers, the dual network reaches the highest and most stable PSNR (solid line in Fig. 14). These observations indicate that convexifying the network into dual forms has a great advantage in making the network converge and stable in a real-world dataset.

3) Verify Zero Dual Gap (Strong Duality): Experiments on CIFAR-10 show that the strong duality between the primal network and the dual network still holds (Fig. 16) when the primal network converges to the global optimality. This observation is further verified by the very close PSNR performance (Table II) and denoised images (Fig. 17) obtained by the two networks.

It should be noted here that the optimal values of the primal ST-CNN and the dual ST-CNN are not exactly equal as the theory proves, but very close (the error is caused by the fact that the experiment is based on a large amount of data training, which is within the negligible range). Hence, experimental results are consistent with our theory.

TABLE III

NAME ABBREVIATIONS OF DIFFERENT METHODS

Abbreviations	Methods
CNN	Two-layer convolutional neural network
Dual_ST_CNN	dual form of CNN with soft-thresholding
Primal_ ST _CNN	CNN with soft-thresholding
ResNet	CNN with ReLU and jump connections

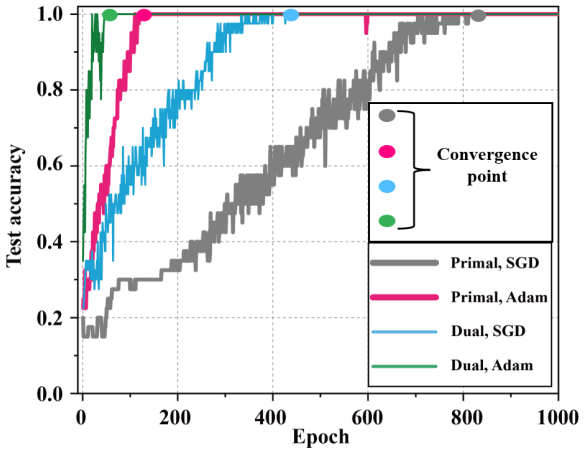


Fig. 18. Classification accuracy when the noise standard deviation is 0.25. Note: For the primal network, reduced accuracy points are observed at the epoch of at the epoch of 595, 596, and 597 when the Adam optimizer is applied.

C. Compare With Other Types of Learning Models

For a more concise description, a simplified table is given (see Table III). CIFAR-10 was used in each experiment. All networks are initialized using a normal distribution with a mean of 0 and a standard deviation of 0.001. The details of the experiment are as follows.

1) Compare the Classification Accuracy With Primal_ST_CNN: In order to realize the classification, classifiers are added to both the Dual_ST_CNN and the Primal_ST_CNN. Fig. 18 shows that the proposed

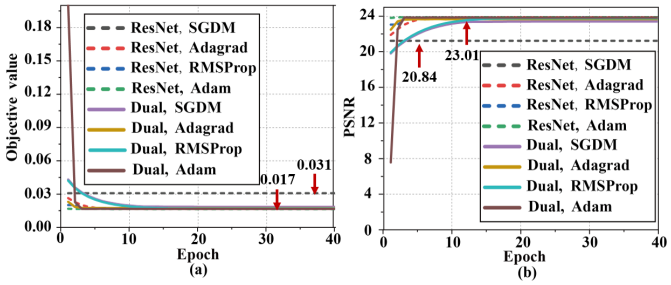


Fig. 19. Comparison between ResNet and our dual ST network. (a) Objective values. (b) PSNR of denoising. Note: The noise standard deviation is 0.25.

Dual_ST_CNN obtains more stable classification accuracy under different epochs than Primal_ST_CNN does. For example, the accuracy of Primal_ST_CNN is reduced to 0.95, 0.975, and 0.975 at the epoch of 595, 596, and 597. This mutation is avoided by Dual_ST_CNN.

2) Compare the Denoising Performance With ResNet: Fig. 19 suggests that the denoising performance of ResNet depends on the optimizers and cannot obtain the optimal performance when SGDM is applied. On the contrary, our Dual_ST_CNN can achieve the best denoising PSNR for all the optimizers.

V. Conclusion

In this article, to achieve the global optimum and remove the dependence of solutions on the initial network parameters, a convex dual convolution neural network with soft-thresholding is proposed to replace its primal convolution neural network with soft-thresholding. Under the principle of convex optimal dual theory, we theoretically analyze the network convexity and prove that the strong duality holds. Extensive results on both simulation and real-world datasets show that strong duality holds, the dual network does not depend on initialization, or optimizer and enables faster convergence than the state-of-the-art two-layer network. This work provides a new way to convexify neural network with soft-thresholding. Furthermore, the convex dual network model of a deep soft-thresholding network with a parallel structure is deduced.

Acknowledgment

The authors would like to thank Jian-Feng Cai, Peng Li, Zi Wang, Yihui Huang, and Nubwimana Rachel for helpful discussions.

References

[1] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[4] Q. Yang, Z. Wang, K. Guo, C. Cai, and X. Qu, "Physics-driven synthetic data learning for biomedical magnetic resonance: The imaging physics-based data synthesis paradigm for artificial intelligence," *IEEE Signal Process. Mag.*, vol. 40, no. 2, pp. 129–140, Mar. 2023.

[5] X. Qu et al., "Accelerated nuclear magnetic resonance spectroscopy with deep learning," *Angew. Chem.*, vol. 132, no. 26, pp. 10383–10386, Jun. 2020.

[6] Y. Huang et al., "Exponential signal reconstruction with deep Hankel matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 6214–6226, Sep. 2023.

[7] Z. Wang et al., "A sparse model-inspired deep thresholding network for exponential signal reconstruction—Application in fast biological spectroscopy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 7578–7592, Oct. 2023.

[8] Y. Ma and D. Klabjan, "Diminishing batch normalization," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9925062

[9] D. Liu, I. W. Tsang, and G. Yang, "A convergence path to deep learning on noisy labels," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9899463

[10] X. Liu, D. Wang, and S.-B. Lin, "Construction of deep ReLU Nets for spatially sparse learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 7746–7760, Oct. 2023.

[11] D. Wang, J. Zeng, and S.-B. Lin, "Random sketching for neural networks with ReLU," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 748–762, Feb. 2021.

[12] S. Krishna Kumar, "On weight initialization in deep neural networks," 2017, *arXiv:1704.08863*.

[13] C. Zhu et al., "Gradinit: Learning to initialize neural networks for stable and efficient training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 16410–16422.

[14] N. Murgovski, L. M. Johannesson, and J. Sjöberg, "Engine on/off control for dimensioning hybrid electric powertrains via convex optimization," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 2949–2962, Sep. 2013.

[15] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Aquatic Invasive. Species*, 2010, pp. 249–256.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[17] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[18] R. Xiong et al., "On layer normalization in the transformer architecture," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 10524–10533.

[19] X. S. Huang, F. Perez, J. Ba, and M. Volkovs, "Improving transformer optimization through better initialization," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 4475–4483.

[20] L. Liu, X. Liu, J. Gao, W. Chen, and J. Han, "Understanding the difficulty of training transformers," 2020, *arXiv:2004.08249*.

[21] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[22] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[24] H. Zhang, Y. N. Dauphin, and T. Ma, "Fixup initialization: Residual learning without normalization," 2019, *arXiv:1901.09321*.

[25] S. De and S. Smith, "Batch normalization biases residual blocks towards the identity function in deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 19964–19975.

[26] A. Brock, S. De, and S. L. Smith, "Characterizing signal propagation to close the performance gap in unnormalized ResNets," 2021, *arXiv:2101.08692*.

[27] A. Brock, S. De, S. L. Smith, and K. Simonyan, "High-performance large-scale image recognition without normalization," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 1059–1071.

[28] Y. Bengio, N. Roux, P. Vincent, O. Delalleau, and P. Marcotte, "Convex neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 123–130.

[29] B. Amos, L. Xu, and J. Z. Kolter, "Input convex neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 146–155.

[30] T. Ergen and M. Pilanci, "Convex geometry of two-layer ReLU networks: Implicit autoencoding and interpretable models," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 4024–4033.

[31] T. Ergen and M. Pilanci, "Convex optimization for shallow neural networks," in *Proc. 57th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2019, pp. 79–83.

[32] M. Pilanci and T. Ergen, "Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 7695–7705.

[33] Y. Wang and M. Pilanci, "The convex geometry of backpropagation: Neural network gradient flows converge to extreme points of the dual convex program," 2021, *arXiv:2110.06488*.

[34] A. Mishkin, A. Sahiner, and M. Pilanci, "Fast convex optimization for two-layer relu networks: Equivalent model classes and cone decompositions," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 15770–15816.

[35] A. Sahiner, T. Ergen, J. Pauly, and M. Pilanci, "Vector-output ReLU neural network problems are copositive programs: Convex analysis of two layer networks and polynomial-time algorithms," 2020, *arXiv:2012.13329*.

[36] T. Ergen and M. Pilanci, "Global optimality beyond two layers: Training deep ReLU networks via convex programs," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 2993–3003.

[37] T. Ergen and M. Pilanci, "Convex geometry and duality of over-parameterized neural networks," *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 9646–9708, 2021.

[38] A. Sahiner, M. Mardani, B. Ozturkler, M. Pilanci, and J. Pauly, "Convex regularization behind neural reconstruction," 2020, *arXiv:2012.05169*.

[39] T. Ergen and M. Pilanci, "Training convolutional relu neural networks in polynomial time: Exact convex optimization formulations," 2020, *arXiv:2006.14798*.

[40] K.-Y. Hsu, H.-Y. Li, and D. Psaltis, "Holographic implementation of a fully connected neural network," *Proc. IEEE*, vol. 78, no. 10, pp. 1637–1645, Oct. 1990.

[41] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 3, pp. 328–339, Mar. 1989.

[42] W. Zhang, J. Tanida, K. Itoh, and Y. Ichioka, "Shift-invariant pattern recognition neural network and its optical architecture," in *Proc. Annu. Conf. Jpn. Soc. Appl. Phys.*, 1988, pp. 2147–2151.

[43] Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[44] C. Cruz, A. Foi, V. Katkovnik, and K. Egiazarian, "Nonlocality-reinforced convolutional neural networks for image denoising," *IEEE Signal Process. Lett.*, vol. 25, no. 8, pp. 1216–1220, Aug. 2018.

[45] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, Jan. 1997.

[46] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[47] P. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 82–90.

[48] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[49] J. Joy, S. Peter, and N. John, "Denoising using soft thresholding," *Int. J. Adv. Res. Elect., Electron. Instrum. Eng.*, vol. 2, no. 3, pp. 1027–1032, 2013.

[50] X.-P. Zhang and M. D. Desai, "Adaptive denoising based on SURE risk," *IEEE Signal Process. Lett.*, vol. 5, no. 10, pp. 265–267, Oct. 1998.

[51] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1532–1546, Sep. 2000.

[52] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[53] M. Zibulevsky and M. Elad, "L1-L2 optimization in signal and image processing," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 76–88, May 2010.

[54] Y. Liu et al., "Projected iterative soft-thresholding algorithm for tight frames in compressed sensing magnetic resonance imaging," *IEEE Trans. Med. Imag.*, vol. 35, no. 9, pp. 2130–2140, Sep. 2016.

[55] X. Zhang et al., "A guaranteed convergence analysis for the projected fast iterative soft-thresholding algorithm in parallel MRI," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101987.

[56] X. Zhang et al., "Accelerated MRI reconstruction with separable and enhanced low-rank Hankel regularization," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2486–2498, Sep. 2022.

[57] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, Jan. 2010.

[58] X. Qu, M. Mayzel, J.-F. Cai, Z. Chen, and V. Orekhov, "Accelerated NMR spectroscopy with low-rank reconstruction," *Angew. Chem. Int. Ed.*, vol. 54, no. 3, pp. 852–854, 2015.

[59] T. Qiu, Z. Wang, H. Liu, D. Guo, and X. Qu, "Review and prospect: NMR spectroscopy denoising and reconstruction with low-rank Hankel matrices and tensors," *Magn. Reson. Chem.*, vol. 59, no. 3, pp. 324–345, 2021.

[60] D. Gnanadurai and V. Sadasivam, "An efficient adaptive thresholding technique for wavelet based image denoising," *Int. J. Signal Process.*, vol. 2, no. 2, pp. 114–119, 2006.

[61] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, "Deep residual shrinkage networks for fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4681–4690, Jul. 2020.

[62] X.-P. Zhang, "Thresholding neural network for adaptive noise reduction," *IEEE Trans. Neural Netw.*, vol. 12, no. 3, pp. 567–584, May 2001.

[63] Z. Wang et al., "One-dimensional deep low-rank and sparse network for accelerated MRI," *IEEE Trans. Med. Imag.*, vol. 42, no. 1, pp. 79–90, Jan. 2023.

[64] T. Lu et al., "PFISTA-SENSE-ResNet for parallel MRI reconstruction," *J. Magn. Reson.*, vol. 318, Sep. 2020, Art. no. 106790.

[65] J.-J. Huang and P. L. Dragotti, "WINNet: Wavelet-inspired invertible network for image denoising," *IEEE Trans. Image Process.*, vol. 31, pp. 4377–4392, 2022.

[66] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[67] M. Sion, "On general minimax theorems," *Pacific J. Math.*, vol. 8, no. 1, pp. 171–176, Mar. 1958.

[68] J. Kindler, "A simple proof of sion's minimax theorem," *Amer. Math. Monthly*, vol. 112, no. 4, pp. 356–358, 2005.

[69] A. Shapiro, "Semi-infinite programming, duality, discretization and optimality conditions," *Optimization*, vol. 58, no. 2, pp. 133–161, 2009.

[70] T. Serra, C. Tjandraatmadja, and S. Ramalingam, "Bounding and counting linear regions of deep neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 4558–4566.

[71] P. C. Ojha, "Enumeration of linear threshold functions from the lattice of hyperplane intersections," *IEEE Trans. Neural Netw.*, vol. 11, no. 4, pp. 839–850, Jul. 2000.

[72] T. M. Cover, "Geometrical and statistical properties of systems of linear Inequal. With applications in pattern recognition," *IEEE Trans. Electron. Comput.*, vol. EC-14, no. 3, pp. 326–334, Jun. 1965.

[73] A. Blum and R. Rivest, "Training a 3-node neural network is NP-complete," in *Adv. Neural Inf. Process. Syst.*, Cambridge, MA, USA, Aug. vol. 5, 1988, pp. 494–501.

[74] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks of the Trade*. New York, NY, USA: Springer, 2012, pp. 437–478.

[75] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—Online stochastic gradient for tensor decomposition," in *Proc. Conf. Learn. Theory*, 2015, pp. 797–842.

[76] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[77] B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro, "Geometry of optimization and implicit regularization in deep learning," 2017, *arXiv:1705.03071*.

[78] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *Proc. 32nd AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1–26.

[79] S. Bhojanapalli et al., "On the reproducibility of neural network predictions," 2021, *arXiv:2102.03349*.

[80] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 1951.

[81] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 315–323.

[82] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[83] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, no. 7, pp. 2121–2159, 2011.

[84] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent," Dept. Comput. Sci., Toronto Univ., Toronto, ON, Canada, Tech. Rep., vol. 14, no. 8, 2012, p. 2. [Online]. Available: https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

[85] X. Chen, S. Liu, R. Sun, and M. Hong, "On the convergence of a class of adam-type algorithms for non-convex optimization," in *Proc. AISTATS*, vol. 84, 2018, pp. 288–297.
[86] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.
[87] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[88] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

Chunyan Xiong received the B.S. degree in mathematics and applied mathematics from Xinyang University, Xinyang, China, in 2017, and the M.S. degree in applied mathematics from Minnan Normal University, Zhangzhou, China, in 2020. She is currently pursuing the Ph.D. degree with the Institute of Electromagnetics and Acoustics, Xiamen University, Xiamen, China.

Her research interests include convex optimization, magnetic resonance imaging, machine learning, image processing, and partial differential equation.

Chaoxing Zhang received the B.E. degree from the School of Electronic Science and Engineering, Xiamen University, Xiamen, China, in 2022, where he is currently pursuing the M.S. degree with the Department of Microelectronics and Integrated Circuit.

His research interests mainly include computer vision, magnetic resonance imaging (cardiac imaging), and he also dabbles in natural language processing.

Mengli Lu received the B.S. degree in computer science from the Jiangxi University of Technology, Ganzhou, China, in 2021. She is currently pursuing the M.S. degree with the Department of Electronic Science, Xiamen University, Xiamen, China.

Her research interests include deep learning, cloud computing, magnetic resonance imaging, and brain region segmentation.

Xiaotong Yu received the B.S. degree in mathematics from Tianjin University, Tianjin, China, in 2021. She is currently pursuing the M.S. degree with the Department of Electronic Science, Xiamen University, Xiamen, China.

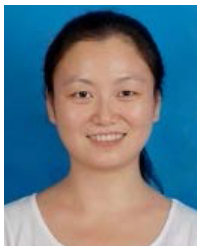
Her research interests include deep learning, cardiac imaging and clinical applications, and magnetic resonance imaging.

Jian Cao received the B.S. degree in mathematics and applied mathematics from the Taiyuan University of Science and Technology, Taiyuan, China, in 2017, and the M.S. degree in applied mathematics from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2020. She is currently pursuing the Ph.D. degree with the Department of Electronic Science, Xiamen University, Xiamen, China.

Her research interests include machine learning, magnetic resonance imaging, and image processing.

Zhong Chen received the B.S. and M.S. degrees in radio physics and the Ph.D. degree in physical chemistry from Xiamen University, Xiamen, China, in 1985, 1988, and 1993, respectively.

He is currently a Distinguished Professor and the Dean of the School of Electronic Science and Engineering, Xiamen University.



Di Guo (Member, IEEE) received the B.S. and Ph.D. degrees in communication engineering from Xiamen University, Xiamen, China, in 2005 and 2012, respectively.

From 2009 to 2011 and from 2018 to 2019, she was a Visiting Scientist with the Department of Electrical Engineering, University of Washington, Seattle, WA, USA. She is currently a Professor with the Department of Computer Science and Technology, Xiamen University of Technology, Xiamen. Her research was supported by the National Natural Science Foundation of China and other grant agencies. Her research interests include computational imaging, magnetic resonance imaging and spectroscopy, signal and image processing, machine learning, artificial intelligence, and cloud computing.

Dr. Guo was a recipient of the IBM Distinguished Student Award in 2012, High-level Talents of Xiamen City in 2019, the First Prize in Natural Science Award of Fujian Province in 2022, and High-level Talents of Fujian Province in 2023.

Xiaobo Qu (Senior Member, IEEE) received the B.S. and Ph.D. degrees in communication engineering from Xiamen University, Xiamen, China, in 2006 and 2011, respectively.

From 2012 to 2019, he was the Visiting Professor with the University of Gothenburg, Gothenburg, Sweden, The Hong Kong University of Science and Technology, Hong Kong, and the University of Washington, Seattle, WA, USA. He is currently a Professor and the Vice Director of the Department of Electronic Science, the Vice Director of Fujian Provincial Key Laboratory of Plasma and Magnetic Resonance, Xiamen University. His research interests include magnetic resonance imaging and spectroscopy, signal and image representations, computational imaging, machine learning, artificial intelligence, and cloud computing.

Dr. Qu has been a senior member of IEEE EMBS and SPS and a member of ISMRM. He was a recipient of the Excellent Young Scientists Fund Award from the National Natural Science Foundation of China in 2021, the First Prize in Natural Science Award of Fujian Province of China in 2022, and the Second Prize of National Teaching Achievement Award in 2023. He is an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and a Senior Editor of *BMC Medical Imaging*.