2C1-OS-06a-5

級数展開に基づく表層非線形ネットワーク

Shallow Nonlinear Network Based on Fourier Series

窪澤駿平 *1*2 Shumpei KUBOSAWA　渡辺太郎 *1 Taro WATANABE　隅田英一郎 *1 Eiichiro SUMITA　岡田将吾 *2 Shogo OKADA　新田克己 *2 Katsumi NITTA

*1情報通信研究機構 先進的音声翻訳研究開発推進センター
ASTREC, National Institute of Information and Communications Technology

*2東京工業大学 大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

One of the most important characteristics of an artificial neural network is the ability of capturing nonlinearity inherent in the dataset. Existing deep architectures take the advantage of capturing global nonlinearity by stacking many hidden layers in order to obtain better representations. We propose an alternative method to handle nonlinearity under a shallow network, i.e., without a deep architecture, by fitting coefficients in a decomposed formula of a multivariate function motivated by Fourier series expansion. Analysis of this network on MNIST datasets reveals that this network enables efficient representation with respect to amount of total parameters. On this network, ℓ_2 regularization greatly helps to avoid overfitting, and L-BFGS method is effective for updating the parameters.

1. はじめに

言語モデルや音声認識，画像認識など様々な識別タスクにおいて，多層パーセプトロン（multi-layered perceptron; MLP）すなわちフィードフォワード・ニューラルネットワーク（feed-forward neural network; FFNN）が広く用いられている．識別タスクにおいて MLP が利用される理由は，並列計算が容易に実現できる単純な計算構造（計算グラフ）によって，データセットに関する線形分離可能性などの性質や，これに基づくカーネルの選び方などの設定によらず，任意の問題を表現可能だからである．任意の問題，すなわちデータセットの入出力関係を表す連続関数は，少なくとも 2 層の MLP によって近似可能であることが証明されている（万能性定理 *1）[1]．

また，近年の深層学習（deep learning）では，多様な識別タスクにおいて，2 層を上回る多層化による汎化性能の向上が示されている．なお，2 層で十分とする万能性定理は，あくまでも「所与である」任意の入出力対に関する表現の万能性を示したものであり，汎化性能その他については触れていない．一方で，深層学習により汎化性能が向上したものの，計算グラフの大規模化により全体のパラメータ数は増大した．また，各層が多数直列に結合されることにより，訓練時の前向き計算の待ち時間が増大し，データ量の増加とは別に計算時間も増加した．

そこで本稿では，少ない層数のニューラルネットワークにおける，汎化性能の向上を目的とした新たなアプローチを提案する．連続関数の大域的な傾向は，関数をフーリエ変換した周波数領域における低周波成分に存在する．そこで本アプローチでは，フーリエ級数展開に基づく計算グラフによってデータセットを表現し，さらに低周波成分を優先的に捉えることで大域的な特徴を獲得する．本アプローチの性質については，MNIST 手書き数字識別タスクを用いた実験結果を基に考察する．

連絡先: NICT　ASTREC　先進的翻訳技術研究室，〒 619-0289 京都府相楽郡精華町光台 3 − 5，e-mail: shumpei.kubosawa@nict.go.jp

*1 一般には入力層も 1 層とみなして「3 層パーセプトロンの表現万能性定理」等呼ばれるが，本稿では活性化関数を含む層のみを数えることとする．

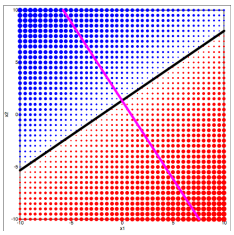

図 1: 2 次元空間における 1 次元へのアフィン変換の効果．格子上の点を変換した際の様子を，各点の色と大きさで表した．色は変換後の符号を，大きさは変換後の値の大小を表す．黒色の直線が識別超平面（変換後の値が 0）であり，これと直交する方向（紫色の直線方向）にのみ識別が有効であることを表す．

2. 多層化の理由

多くの識別タスクにおける入出力は実数ベクトルである．このため，入出力の次元数をそれぞれ $d_i, d_o \in \mathbb{N}$ とすると，識別タスクは，データセットの入力データ空間からラベル空間への写像 $f : \mathbb{R}^{d_i} \to \mathbb{R}^{d_o}$ を得る学習タスクとして定式化される．MLP を含むニューラルネットワークは一般に，予めパラメータと計算順序および入出力により定義される計算グラフ（関数定義）を用意し，そのパラメータを訓練データセットにフィッティングすることにより識別器を得る枠組みである．MLP は，複数のパーセプトロンを入力側から出力側へ直列に接続した識別モデルである．一方で，1 層のパーセプトロンであっても，識別可能な問題は多く存在する．例えば，シグモイド関数 $\sigma(x) = \frac{1}{1+e^{-x}}$ を活性化関数とするパーセプトロンはロジスティック回帰と等価である．では，なぜ多層化が必要なのだろうか？

多層化の理由の第一は，線形分離不可能な問題を含む任意の問題を表現するためである．Minsky[5] は，パーセプトロンでは線形分離不可能な問題の表現と訓練の両方が不可能であることを示した．そもそも，MLP の各層を構成するパーセプトロンは，入出力の次元数を $l_i, l_o \in \mathbb{N}$ とすると，入力ベク

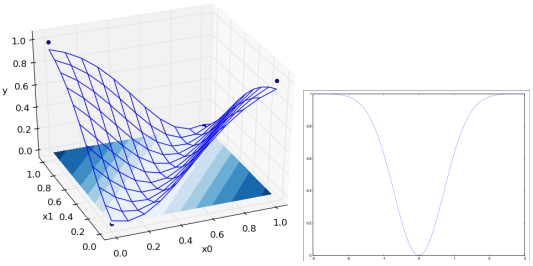

図 2: flatabs 関数を活性化関数とした場合の 1 層パーセプトロンが XOR 問題を識別する様子．左：学習結果（メッシュ）と教師信号点（小球）．右：flatabs 関数の形状．

トル $\mathbf{x} \in \mathbb{R}^{l_i}$ について $\mathbf{y} = \sigma(W\mathbf{x} + \mathbf{b})$ を出力する．ここで，$W \in \mathbb{R}^{l_o} \times \mathbb{R}^{l_i}$ は重み行列，$\mathbf{b} \in \mathbb{R}^{l_o}$ はバイアスベクトル，$\sigma : \mathbb{R} \to \mathbb{R}$ はベクトルの要素毎に適用される活性化関数であり，フィッティング対象のパラメータは W と $\mathbf{b}$ である．線形分離可能性という制約は，このパーセプトロンの計算過程における 2 要素に起因している．1 個目はアフィン変換 $W\mathbf{x} + \mathbf{b}$ であり，2 個目は（広義）単調な活性化関数 σ である．アフィン変換は，入力空間上の各データ点を，ある 1 直線（あるいは超平面）からの符号付きの距離に変換する．アフィン変換の効果を図 1 に示す．活性化関数 σ は，入力ベクトルにアフィン変換を適用したベクトル $W\mathbf{x} + \mathbf{b}$ の各要素に適用される．つまり，アフィン変換後に単調関数を適用する構造のため，いかなる超平面でも二分できない場合は識別不可能である．一方で，活性化関数を非単調な関数，例えば $\sigma(x) = 1 - e^{-x^2}$（以後，flatabs 関数と呼ぶ）に置き換えると，線形分離不可能な XOR 問題であっても，1 層パーセプトロンによって識別可能である．この様子を図 2 に示す．ただし，識別性能及び訓練効率について最適な活性化関数を，事前に設計することは困難である．そこで，少なくとも 1 層パーセプトロンの各素子には入力空間を超平面で二分する識別能力があることを利用し，まず 1 層目（隠れ層）で入力空間を様々な超平面で二分し，次の層（出力層）ではさらにその層の入力空間を様々な超平面で二分する，すなわち 1 層目で様々に二分された部分入力空間の組み合わせによって，任意の関数を表現可能にしたものが 2 層 MLP である．2 層 MLP が任意の問題を表現可能であることは万能性定理によって裏付けられている．一方で，本質的には 2 層のパーセプトロンで十分のはずが，深層学習ではそれを上回る層数が用いられている．なぜなのだろうか？

多層化の理由の第二は，過適合の抑制である．2 層パーセプトロンの場合，訓練データセットの識別誤差を減らすために 1 層目の素子数を増やすと，過適合する傾向がある．これは 1 層目で入力空間を様々に二分する際に，各訓練データ点を個別に表現する様に超平面が構成され，訓練データが存在しない領域における補間作用が働かなくなるためである．一方，深層学習では，入力側の層から出力側の層に向かって，入力空間を様々に二分（1 層目）して出来る部分入力空間の組み合わせ（2 層目）により出来る部分入力空間の組み合わせ（3 層目）…という様に，組み合わせによる表現が強制される．このため，個々のデータ点という局所への最適化が起こりづらくなる．つまり，接続された各 2 層は，十分な素子数があれば万能な表現が可能だが，「各層における部分入力空間の組み合わせで訓練データ全体を表現せよ」という制約がネットワーク全体に与えられることで，訓練データセットの大域的な傾向を捉えやすくなり，汎化性能が向上するものと考えられる．

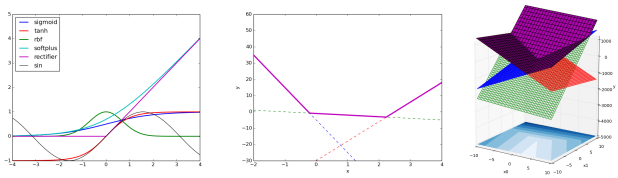

図 3: 様々な活性化関数の形状．左：図中左上の凡例に示す各関数の形状．中央・右：引数がスカラーおよび 2 次元ベクトルの場合の maxout 関数形状（紫色）の例．

3. 関連研究

誤差関数の勾配と誤差逆伝播に基づいて学習されるニューラルネットワークにおいて，活性化関数に求められる条件は（劣）微分可能性である．そこで，この条件を満たす様々な関数が活性化関数として使用されてきた．これまで取り上げてきたシグモイド関数に加え，ソフトマックス関数やハイパーボリックタンジェントなどは，勾配計算が容易であることから広く用いられている．他にも，(広義）単調関数としては softplus $\sigma(x) = \log(1 + e^x)$ や ReLU $\sigma(x) = \max(0, x)$（rectified linear unit）[2] などが提案されており，非単調関数としては放射基底関数（radial basis function; RBF）や maxout[3]，さらに sin，cos（以後，これらをまとめてシヌソイドと呼ぶ）などが使用されてきた．図 3 左に，スカラーを入力としたときのこれらの関数の形状を示す．ソフトマックス関数は，基本的にシグモイド関数と同様の形状である．

ReLU は，softplus を区分線形化したものであり，ℓ_1 正則化と併用することでスパースな内部表現を得やすい特徴がある．ReLU と ℓ_1 正則化を組み合わせると，値が 0 となるパラメータが増えることが期待される．これは必要最低限度の素子数で識別を行う制約として働くため，汎化性能が向上したものと考えられる．一方で，ReLU は多層化を前提としており，また機能しなくなる素子の発生を見込んでネットワークを設計する必要があり，パラメータ数は増加する．

RBF ネットワークは，2 層 MLP の隠れ層における活性化関数が放射基底関数であるものを指し，これも任意の連続関数を近似可能であることが示されている [7]．一方で，個々の RBF 素子は入力空間の局所的な特徴を捉えるため，隠れ層の素子数を増やすことにより訓練データでの識別性能は向上するが，汎化性能の向上については必ずしも期待できない．これは 2 層パーセプトロンと同様である．加えて，訓練データが複雑な分布である場合，素子数の増加が避けられない．

maxout は，$\sigma(\mathbf{x}) = \max(\mathbf{w}_1^\top \mathbf{x} + b_1, \mathbf{w}_2^\top \mathbf{x} + b_2, \cdots, \mathbf{w}_m^\top \mathbf{x} + b_m)$ ただし $\mathbf{w}_\star \in \mathbb{R}^{l_i}$，$b_\star \in \mathbb{R}$，$m \in \mathbb{N}$ という関数であり，パラメータは $\mathbf{w}_\star$ と $b_\star$ である．つまり，各 maxout 素子の内部には m 個のサブモデル（アフィン変換）があり，それらの最大値を関数出力として採用するものである．maxout の入力が 1 次元の場合と 2 次元の場合の関数形状を図 3 中央および右に示す．maxout は，個々のサブモデルの最大値を用いることで，素子の入力空間における区分線形化された凸関数を構成する．これも，入力空間の局所的な特徴を捉える関数であるため，本質的に多層化を前提とした活性化関数である．

シヌソイドを活性化関数として用いる 2 層 MLP は，フーリエ級数展開による関数近似と等価であると捉えられる．例えば，隠れ層の活性化関数が sin である場合，このネットワークの n 番目の出力は次式により表される：

$$y_n = \mathbf{a}_n^\top \sin(W\mathbf{x} + \mathbf{b})$$

ただし $\mathbf{a}_\star \in \mathbb{R}^h$，$h \in \mathbb{N}$（隠れ層の sin 素子数）であり，出力層の活性化関数は恒等写像としている（出力層のバイアスも省略）．これは，フーリエ正弦級数展開と等価であるため，任意の奇関数を表現可能である．ある訓練データセットについてパラメータ $\mathbf{a}_\star$，W，$\mathbf{b}$ を最適化した場合，$\mathbf{a}_\star$，W，$\mathbf{b}$ は，訓練データセット全体の周波数領域における振幅，周波数，位相に対応する．なお，sin と cos の違いは位相のみだが，位相もパラメータであるため，仮に sin または cos のどちらか一方のみを活性化関数として採用したとしても，任意の連続関数を表現可能である．

4. 表層非線形ネットワーク

少ない層数のニューラルネットワークによって汎化性能の向上を図るためには，データセットの大域的な特徴を効率的に表現する必要がある．そこで本稿では，フーリエ級数展開に基づいているシヌソイドを活性化関数として用いる 2 層 MLP に関連して，次に示す 3 個の手法を組み合わせたアプローチを提案する：

1. 隠れ層の各出力は，各出力に個別のフーリエ級数展開表現を用いる．
2. パラメータ最適化に L-BFGS 法 [6] を用いる．
3. 周波数パラメータ最適化の際に ℓ_2 正則化を行う．

提案手法で用いるネットワークの計算グラフを図 4 に示す．深層学習の場合は，入力に近い層から順に，画像で言うエッジなど学習対象の概念を構成する基本部品（局所的な分布）が学習され，層が出力に近づくにつれて学習対象の概念（大域的な分布）が構成的に学習される，と説明される．一方，本アプローチのネットワークはフーリエ展開に基づき，入力について複数の重み付けをされた cos 素子が並列していることにより，訓練データセットにおける局所的な分布は cos の入力の時点で大きく重みづけられ（高周波成分），大域的な分布は同様に小さく重みづけられること（低周波成分）で表現される．また，2 層目が全結合ではないことにより，2 層目の各素子出力（図中 hidden unit）は個別の非線形関数として機能する．以後，この非線形関数の単位を非線形ユニットと呼ぶ．非線形ユニットは，識別器という観点からは特徴量化を担うカーネルの様な役割を持つ．従来のシヌソイドを用いた 2 層 MLP では出力層も全結合であったが，これを排することによって，非線形ユニットがそれぞれ個別に最適化される設計とした．

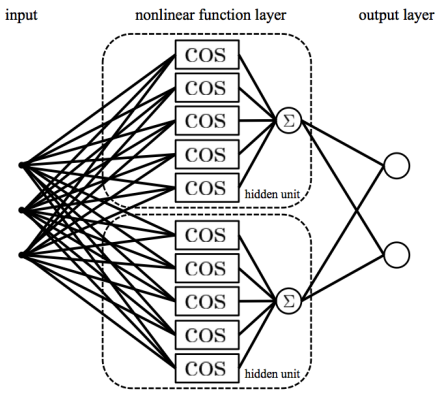

図 4: 本アプローチで用いるネットワーク構造（計算グラフ）

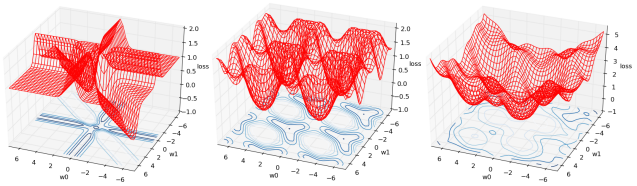

図 5: 活性化関数の周期性と ℓ_2 正則化が，誤差関数の形状に与える影響の例．左：flatabs 関数による活性化．中央：cos 活性化・ℓ_2 正則化なし．右：cos 活性化・ℓ_2 正則化あり．

パラメータ最適化については，本アプローチも誤差逆伝播に基づくが，ニューラルネットワークで一般に用いられる最急降下法は用いず，L-BFGS 法を用いる．誤差関数は従来のネットワークと同じ 2 乗誤差またはクロスエントロピー誤差を用いるが，シヌソイドを活性化関数として用いるネットワークでは誤差関数の形状が問題となる．1 層パーセプトロンにおける XOR 問題を例に，活性化関数として flatabs 関数を用いた場合と cos を用いた場合の重み行列に関する 2 乗誤差関数の形状を図 5（左，中央）に示す（ただし，cos は値域を $[0, 1]$ とするため活性化関数を $\sigma(x) = \frac{\cos(x)+1}{2}$ とし，バイアスはそれぞれ適当な値に固定した）．cos の周期性が誤差関数にも現れることが図に示されている．この例では，どの局所最小値であっても同じ誤差値が得られるが，より複雑な問題では至る所に極小値が存在する．このため，誤差関数の現在のパラメータ点における最急勾配方向へパラメータを移動させる最急降下法では，妥当な解に収束することが期待できない．一方で，この凹凸が激しくなるのは cos の引数側のパラメータに関してのみである．そこで，前向き計算における cos より前のパラメータ，すなわち周波数パラメータに ℓ_2 正則化項を加えることで，ある程度誤差関数を滑らかにする．ℓ_2 正則化項を追加した場合の誤差関数の形状を図 5（右）に示す．ℓ_2 正則化の影響により，誤差関数は大域的には凸となる．ここで準ニュートン法である L-BFGS 法を用いることで，妥当な解への収束が期待される．

周波数パラメータの ℓ_2 正則化には，収束させること以外にも重要な目的がある．画像を例にすると，自然画像をフーリエ変換した場合，大域的な情報は低周波領域に存在する．高周波成分は，主にエッジや孤立した点などの，極端に値が変化する箇所の存在を表している．一般のデータセットにおけるこれら高周波成分は，外れ値の存在を表すものと考えられる．一方で，本アプローチでフーリエ級数を利用するのは，大域的な特徴を効率よく捉えることが目的である．そこで，周波数パラメータについて ℓ_2 正則化することにより，低周波領域でデータセット全体を近似する．画像処理における標準画像を 2 値化したものを例に，周波数パラメータの ℓ_2 正則化の効果を図 6 に示す．ここでは画像を，各ピクセルの位置座標が入力データであり，各座標の輝度が教師信号であるデータセットとみなしている．ℓ_2 正則化を加えない場合は局所的な特徴を捉えたことによるノイズが見られるが，ℓ_2 正則化を行った場合は大域的な特徴のみを獲得していることが，図 6 に示されている．

5. 実験結果

提案手法による識別タスクでの汎化性能とパラメータ数について，MNIST 手書き数字データを用いて，maxout および ReLU（rectifier）と比較した．MNIST は，入力が 28×28 画素のグレースケール画像の各行を並べた 784 次元ベクトルであり，入力ベクトルを数字 10 クラスに分類するタスクである．

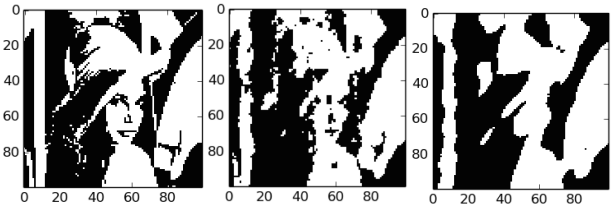

図 6: 画像をデータセットとして用いて可視化した周波数パラメータへの ℓ_2 正則化の効果．左：元画像（教師信号）．中央：ℓ_2 正則化なしの学習結果．右：ℓ_2 正則化ありの学習結果．

全 70,000 データのうち，訓練データセットは 60,000 データであり，テストデータセットは残りの 10,000 データである．本アプローチでは事前学習（pre-training）を行わないため，事前学習を行わないモデルのみを比較対象として引用する．比較対象のモデルにおけるパラメータ数は，各論文および公開されているソースコードより算出した．全ての手法において，出力層の活性化関数はソフトマックス関数である．この設定における maxout は全 3 層のモデルであり，rectifier は全 4 層である．また，maxout では正則化に dropout が用いられている．テストデータセットにおける識別誤り率の比較を表 1 に示す．表中，提案手法の名称の後の括弧内の数値は，順に非線形ユニット数と各ユニット内の cos 素子数を表す．本手法は，従来提案されてきた手法と比べて，汎化性能は及ばなかった．一方で，識別性能に対するパラメータ数の観点からは，各パラメータが他のモデルよりも識別に大きく寄与していることが裏付けられた．

表 1: MNIST データセットを用いた識別誤り率の比較

識別モデル	識別誤り率	パラメータ数
maxout[3]	0.94%	1,233K
rectifier[2]	1.43%	3,798K
提案手法（30/10）	1.79%	472K
提案手法（20/10）	1.81%	157K
提案手法（10/15）	1.84%	118K

6. 考察

ここで，提案手法が汎化性能の点で従来手法に及ばなかったことについて考察する．提案手法では，周波数パラメータの ℓ_2 正則化により汎化性能を得ようとしている．そこで，ℓ_2 正則化を行う場合と行わない場合とを比較する．これらの 2 条件における識別誤り率とイテレーション回数の関係を図 7 に示す．この実験では，MNIST の 50,000 件を訓練データセットとし，残りの 10,000 件ずつをバリデーションデータセットとテストデータとして評価に用いた．図より，ℓ_2 正則化は確かに効果的だが，学習の中盤（400 イテレーション）以降で ℓ_2 正則化だけでは抑制しきれない過学習が発生していることが判明した．また，学習の進行に伴い，出力層の重み行列と振幅パラメータのノルムが，訓練データの識別誤り率にのみ関係して増加する現象を確認している．このため，これらのパラメータについて ℓ_2 正則化やノルム正規化 [4] を行ったが，汎化性能は向上しなかった．従って，振幅パラメータあるいは周波数パラメータの分布に関する他の制約や，周波数領域におけるサンプリング等の方法を検討する余地があると考察される．

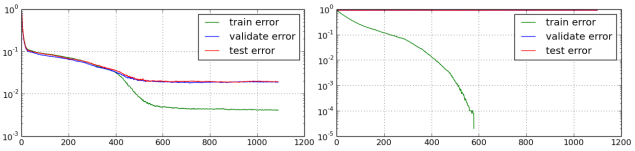

図 7: 訓練時のイテレーション回数（横軸）と識別誤り率（縦軸・対数目盛）の関係に対する，周波数パラメータへの ℓ_2 正則化が与える影響．左：ℓ_2 正則化あり．右：ℓ_2 正則化なし．ℓ_2 正則化を行わないと汎化作用がほぼ無いことが示されている．

7. おわりに

本稿では，少ない層数のネットワーク構造により，パラメータ数を減らして前向き計算の効率を上げることと，汎化性能の向上を目的とした，ニューラルネットワークの新たなアプローチを提案した．提案手法のネットワークはフーリエ級数展開に基づく構造であり，データセットをフーリエ変換した際の低周波成分で近似するという制約を与えることで，データセットの大域的な特徴を得て汎化性能を上げるアプローチをとった．MNIST データセットによる評価実験では，汎化性能は従来手法を上回らなかったが，パラメータ数という観点からは効率的に識別されていることを確認した．一方で，低周波成分による近似だけでは，従来手法による汎化性能を上回ることが難しいことが判明した．このため，汎化性能向上のために更なる制約を追加する等の方法について，検討を進める予定である．

参考文献

[1] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems, 2(4), 303-314.

[2] Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier networks. Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP Volume. Vol. 15.

[3] Goodfellow, I., Warde-farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013). Maxout Networks. In Proceedings of the 30th International Conference on Machine Learning (ICML-13), pages 1319-1327.

[4] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.

[5] Minsky, M., Papert, S. (1972). Perceptrons: An Introduction to Computational Geometry, The MIT Press, Cambridge MA.

[6] Liu, D. C., and Nocedal, J. (1989). ”On the Limited Memory Method for Large Scale Optimization”. Mathematical Programming B 45 (3): 503–528. doi:10.1007/BF01589116.

[7] Park, J., and Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks. Neural computation, 3(2), 246-257.