

Deep Neural Networks and Variational Inequalities

Jean-Christophe Pesquet

Center for Visual Computing, CentraleSupélec, University Paris-Saclay-Inria

Joint work with Patrick Louis Combettes

Workshop on Operator Splitting Methods in Data Analysis - March 2019

Motivation

Background

NEURAL NETWORKS

- ✓ Generic methods for nonlinear approximation
[Cybenko, 1989]
- ✓ Efficient for incorporating prior knowledge from big databases
- ✗ Black-box, empirical approaches

Background

NEURAL NETWORKS

- ✓ Generic methods for nonlinear approximation
[Cybenko, 1989]
- ✓ Efficient for incorporating prior knowledge from big databases
- ✗ Black-box, empirical approaches

Objective

Show that optimization concepts are not only useful to train NNs, but also to analyze them

Feedforward NNs



NEURAL NETWORK MODEL

Let $m \geq 1$ be an integer and let $(H_i)_{0 \leq i \leq m}$ be nonzero real Hilbert spaces.

$$T = T_m \circ \cdots \circ T_1$$

where $(\forall i \in \{1, \dots, m\}) \quad T_i: H_{i-1} \rightarrow H_i: x \mapsto R_i(W_i x + b_i)$,

$W_i: H_{i-1} \rightarrow H_i$ is a bounded linear (weight), operator

b_i is a (bias) vector in H_i ,

and $R_i: H_i \rightarrow H_i$ is an α_i -averaged (activation) operator with $\alpha_i \in [0, 1]$, i.e. there exists a nonexpansive operator Q_i such that

$$R_i = (1 - \alpha_i)\text{Id} + \alpha_i Q_i.$$

NNs and prox

Standard activation operators

Most of them are proximity operators (1/2-averaged)

- Rectified linear unit (ReLU)

$$\varrho: \mathbb{R} \rightarrow \mathbb{R}: \xi \mapsto \begin{cases} \xi, & \text{if } \xi > 0; \\ 0, & \text{if } \xi \leq 0. \end{cases}$$

Then, $\varrho = \text{proj}_{[0, +\infty[} \cdot$

- Parametric ReLU

$$\varrho: \mathbb{R} \rightarrow \mathbb{R}: \xi \mapsto \begin{cases} \xi, & \text{if } \xi > 0; \\ \alpha \xi, & \text{if } \xi \leq 0 \end{cases}, \quad \alpha \in]0, 1].$$

Then $\varrho = \text{prox}_{\phi}$ where

$$\phi: \mathbb{R} \rightarrow \mathbb{R}: \xi \mapsto \begin{cases} 0, & \text{if } \xi > 0; \\ (1/\alpha - 1)\xi^2/2, & \text{if } \xi \leq 0. \end{cases}$$

Standard activation operators

Most of them are proximity operators (1/2-averaged)

- Unimodal sigmoid

$$\varrho: \mathbb{R} \rightarrow \mathbb{R}: \xi \mapsto \frac{1}{1 + e^{-\xi}} - \frac{1}{2}$$

Then $\varrho = \text{prox}_\phi$ where

$$\phi: \xi \mapsto \begin{cases} (\xi + 1/2) \ln(\xi + 1/2) + (1/2 - \xi) \ln(1/2 - \xi) - \frac{1}{2}(\xi^2 + 1/4) & \text{if } |\xi| < 1/2; \\ -1/4, & \text{if } |\xi| = 1/2; \\ +\infty, & \text{if } |\xi| > 1/2. \end{cases}$$

- Elliot function

$$\varrho: \mathbb{R} \rightarrow \mathbb{R}: \xi \mapsto \frac{\xi}{1 + |\xi|}.$$

We have $\varrho = \text{prox}_\phi$, where

$$\phi: \mathbb{R} \rightarrow]-\infty, +\infty]: \xi \mapsto \begin{cases} -|\xi| - \ln(1 - |\xi|) - \frac{\xi^2}{2}, & \text{if } |\xi| < 1; \\ +\infty, & \text{if } |\xi| \geq 1. \end{cases}$$

Standard activation operators

Most of them are proximity operators (1/2-averaged)

- Softmax

$$R: \mathbb{R}^N \rightarrow \mathbb{R}^N: (\xi_k)_{1 \leq k \leq N} \mapsto \left(\exp(\xi_k) / \sum_{j=1}^N \exp(\xi_j) \right)_{1 \leq k \leq N} - u,$$

where $u = (1, \dots, 1)/N \in \mathbb{R}^N$.

Then $R = \text{prox}_\varphi$ where $\varphi = \psi(\cdot + u) + \langle \cdot | u \rangle$ and

$$\psi: \mathbb{R}^N \rightarrow]-\infty, +\infty]$$

$$(\xi_k)_{1 \leq k \leq N} \mapsto \begin{cases} \sum_{k=1}^N \left(\xi_k \ln \xi_k - \frac{\xi_k^2}{2} \right), & \text{if } (\xi_k)_{1 \leq k \leq N} \in [0, 1]^N \\ & \text{and } \sum_{k=1}^N \xi_k = 1; \\ +\infty, & \text{otherwise.} \end{cases}$$

Standard activation operators

Most of them are proximity operators (1/2-averaged)

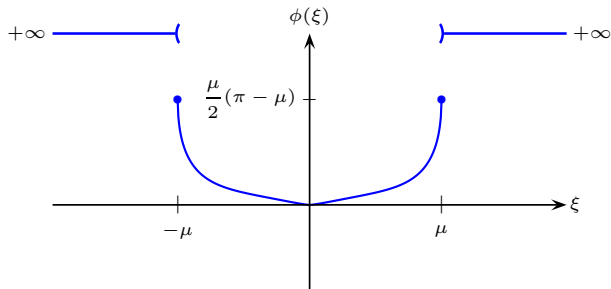
- Squashing function used in capsnets

$$(\forall x \in \mathbb{R}^N) \quad Rx = \frac{\mu \|x\|}{1 + \|x\|^2} x = \text{prox}_{\phi \circ \|\cdot\|} x, \quad \mu = \frac{8}{3\sqrt{3}},$$

where

$$\phi: \xi \mapsto \begin{cases} \mu \arctan \sqrt{\frac{|\xi|}{\mu - |\xi|}} - \sqrt{|\xi|(\mu - |\xi|)} - \frac{\xi^2}{2}, & \text{if } |\xi| < \mu; \\ \frac{\mu(\pi - \mu)}{2}, & \text{if } |\xi| = \mu; \\ +\infty, & \text{otherwise.} \end{cases}$$

Standard activation operators



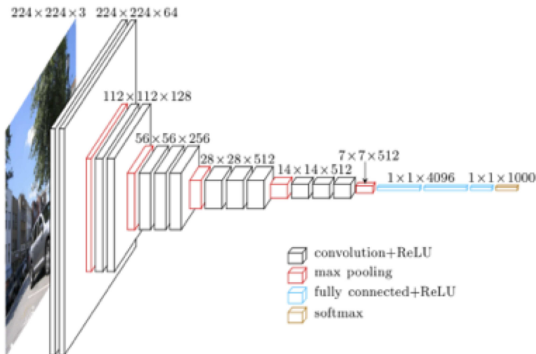
Characterization of NNs

Averageness result

GOAL

Characterize properties of compositions of linear operators and firmly nonexpansive mappings

DIFFICULTY



Averageness result

[Combettes, Pesquet, 2018]

Let $\alpha \in [1/2, 1]$. Let $W = W_m \circ \dots \circ W_1$, let $\mu = \inf_{x \in H_0, \|x\|=1} \langle Wx \mid x \rangle$, and let

$$\theta_m = \|W\| + \sum_{\ell=1}^{m-1} \sum_{0 \leq j_1 < \dots < j_\ell \leq m-1} \|W_m \circ \dots \circ W_{j_\ell+1}\| \|W_{j_\ell} \circ \dots \circ W_{j_\ell-1+1}\| \dots \|W_{j_1} \circ \dots \circ W_0\|.$$

If one of the following conditions is satisfied:

- (i) $\|W - 2^m(1 - \alpha)\text{Id}\| - \|W\| + 2\theta_m \leq 2^m\alpha$;
- (ii) $\alpha \neq 1$, for every $k \in \{1, \dots, m\}$ $W_k \neq 0$, and there exists $\eta \in [0, \alpha/((1 - \alpha)\theta_m)]$ such that

$$\begin{cases} \theta_m \leq 2^m\alpha \\ \alpha\theta_m + (1 - \alpha)(\|\text{Id} - \eta W\| - \eta\|W\|)(\theta_m - \|W\|) \leq 2^{m-1}(2\alpha - 1) + (1 - \alpha)\mu, \end{cases}$$

then $T = R_m \circ (W_m \cdot + b_m) \circ \dots \circ R_0 \circ (W_0 \cdot + b_0)$ is α -averaged.

Open issue 1: more general conditions

POSITIVE RESULTS

$\theta_m/2^{m-1}$ is a **Lipschitz constant** of T [Combettes, Pesquet, 2019]

When $\alpha = 1$, conditions (i) or (ii) reduces to

$$\theta_m/2^{m-1} \leq 1.$$

Necessary condition: $\|W_m \circ \dots \circ W_1\| \leq 1$.

Sufficient condition: $\prod_{i=1}^m \|W_i\| \leq 1$.

↪ Less restrictive than imposing

$$(\forall i \in \{1, \dots, m\}) \quad \|W_i\| \leq 1.$$

Open issue 1: more general conditions

NEGATIVE RESULTS

Case when $H_1 = \dots = H_m$ and $W_1 = \dots = W_m = \text{Id}$.

- Condition (i): $\alpha = 1 - \frac{1}{2^m}$
- Condition (ii): $\alpha = \frac{2^{m-1}}{1+2^{m-1}}$
- [Combettes, Yamada, 2015]: $\alpha = \frac{m}{m+1}$

Open issue 1: more general conditions

For every $i \in \{1, \dots, m\}$, set

$$L_i: H_0 \times \dots \times H_{i-1} \rightarrow H_i: (x_k)_{0 \leq k \leq i-1} \mapsto \sum_{k=0}^{i-1} (W_i \circ \dots \circ W_{k+1})x_k.$$

For every $\mathbf{x} = (x_i)_{0 \leq i \leq m-1} \in H_0 \times \dots \times H_{m-1}$ such that

$$(\forall i \in \{0, \dots, m-1\}) \quad \|x_i\| \leq \begin{cases} 1, & \text{if } i = 0; \\ \|L_i(x_0, \dots, x_{i-1})\|, & \text{if } i \geq 1, \end{cases}$$

then

$$\|L_m \mathbf{x} - 2^m(1 - \alpha)x_0\| + \|L_m \mathbf{x}\| \leq 2^m \alpha \|x_0\|.$$

Asymptotic behavior

Investigated model

MODEL

Let $x_0 \in H$ and let $\{\lambda_n\}_{n \in \mathbb{N}} \subset]0, +\infty[$. Iterate

for $n = 0, 1, \dots$

$$\left[\begin{array}{l} x_{1,n} = R_{1,n}(W_{1,n}x_n + b_{1,n}) \\ x_{2,n} = R_{2,n}(W_{2,n}x_{1,n} + b_{2,n}) \\ \vdots \\ x_{m,n} = R_{m,n}(W_{m,n}x_{m-1,n} + b_{m,n}) \\ x_{n+1} = x_n + \lambda_n(x_{m,n} - x_n). \end{array} \right.$$

- $(\forall n \in \mathbb{N})(\forall i \in \{1, \dots, m\})$ $W_{i,n}: H_{i-1} \rightarrow H_i$ is a bounded linear operator, $b_{i,n} \in H_i$, and $R_{i,n}: H_i \rightarrow H_i$
- $(H_i)_{0 \leq i \leq m}$ nonzero real Hilbert spaces such that $H_m = H_0 = H$.

Investigated model

MODEL

Let $x_0 \in H$ and let $\{\lambda_n\}_{n \in \mathbb{N}} \subset]0, +\infty[$. Iterate

for $n = 0, 1, \dots$

$$\left\{ \begin{array}{l} x_{1,n} = R_{1,n}(W_{1,n}x_n + b_{1,n}) \\ x_{2,n} = R_{2,n}(W_{2,n}x_{1,n} + b_{2,n}) \\ \vdots \\ x_{m,n} = R_{m,n}(W_{m,n}x_{m-1,n} + b_{m,n}) \\ x_{n+1} = x_n + \lambda_n(x_{m,n} - x_n). \end{array} \right.$$

REMARK

$(\lambda_n)_{n \in \mathbb{N}} \sim \text{skip connections}$

Static analysis

ASSUMPTIONS

- $R_{i,n} \equiv R_i, W_{i,n} \equiv W_i, b_{i,n} \equiv b_i$
- $(\forall i \in \{1, \dots, m\}) R_i = \text{prox}_{\varphi_i}$ for some $\varphi_i \in \Gamma_0(H_i)$ with $\varphi_i(0) = \inf \varphi_i(H_i)$.

NOTATION

- For every $i \in \{1, \dots, m\}$, $T_i: H_{i-1} \rightarrow H_i: x \mapsto R_i(W_i x + b_i)$
- $F = \text{Fix } T$ with $T = T_m \circ \dots \circ T_1$.

Static analysis

VARIATIONAL INEQUALITY PROBLEM

Find $\bar{x}_1 \in H_1, \dots, \bar{x}_m \in H_m$ such that

$$\begin{cases} b_1 \in \bar{x}_1 - W_1 \bar{x}_m + \partial\varphi_1(\bar{x}_1) \\ b_2 \in \bar{x}_2 - W_2 \bar{x}_1 + \partial\varphi_2(\bar{x}_2) \\ \vdots \\ b_m \in \bar{x}_m - W_m \bar{x}_{m-1} + \partial\varphi_m(\bar{x}_m). \end{cases} \quad (1)$$

- The set of solutions to (1) is
 $F = \{(T_1 \bar{x}_m, (T_2 \circ T_1) \bar{x}_m, \dots, (T_{m-1} \circ \dots \circ T_1) \bar{x}_m, \bar{x}_m) \mid \bar{x}_m \in F\}.$
- Suppose that $(W_i)_{1 \leq i \leq m}$ satisfies averageness conditions for some $\alpha \in [1/2, 1]$. Then F is closed and convex.
- In addition, if one of the following holds:
 - (a) $\text{ran}(T_m \circ \dots \circ T_1)$ is bounded.
 - (b) There exists $j \in \{1, \dots, m\}$ such that $\text{dom } \varphi_j$ is bounded.

Then F and F are nonempty.

Static analysis

VARIATIONAL INEQUALITY PROBLEM

Find $\bar{x}_1 \in H_1, \dots, \bar{x}_m \in H_m$ such that

$$\begin{cases} b_1 \in \bar{x}_1 - W_1 \bar{x}_m + \partial \varphi_1(\bar{x}_1) \\ b_2 \in \bar{x}_2 - W_2 \bar{x}_1 + \partial \varphi_2(\bar{x}_2) \\ \vdots \\ b_m \in \bar{x}_m - W_m \bar{x}_{m-1} + \partial \varphi_m(\bar{x}_m). \end{cases} \quad (1)$$

Remark: A solution to (1) usually does not solve a minimization problem.

Asymptotic analysis

Suppose that the following hold:

- (a) $F \neq \emptyset$.
- (b) $(W_i)_{1 \leq i \leq m}$ satisfies averageness conditions with parameter α .
- (c) One of the following is satisfied:
 - (i) $\lambda_n \equiv 1/\alpha = 1$ and $Tx_n - x_n \rightarrow 0$.
 - (ii) $(\lambda_n)_{n \in \mathbb{N}}$ lies in $]0, 1/\alpha[$ and $\sum_{n \in \mathbb{N}} \lambda_n (1 - \alpha \lambda_n) = +\infty$.

Then $(x_n)_{n \in \mathbb{N}}$ converges weakly to a point $\bar{x}_m \in F$ and $(T_1 \bar{x}_m, (T_2 \circ T_1) \bar{x}_m, \dots, (T_{m-1} \circ \dots \circ T_1) \bar{x}_m, \bar{x}_m)$ solves (1).

Asymptotic analysis

Suppose that the following hold:

- (a) $F \neq \emptyset$.
- (b) $(W_i)_{1 \leq i \leq m}$ satisfies averageness conditions with parameter α .
- (c) One of the following is satisfied:
 - (i) $\lambda_n \equiv 1/\alpha = 1$ and $Tx_n - x_n \rightarrow 0$.
 - (ii) $(\lambda_n)_{n \in \mathbb{N}}$ lies in $]0, 1/\alpha[$ and $\sum_{n \in \mathbb{N}} \lambda_n (1 - \alpha \lambda_n) = +\infty$.

Then $(x_n)_{n \in \mathbb{N}}$ converges weakly to a point $\bar{x}_m \in F$ and $(T_1 \bar{x}_m, (T_2 \circ T_1) \bar{x}_m, \dots, (T_{m-1} \circ \dots \circ T_1) \bar{x}_m, \bar{x}_m)$ solves (1).

Remark: Suppose that $(W_i)_{1 \leq i \leq m}$ satisfies averageness conditions with $\alpha \in [1/2, 1]$, $(\lambda_n)_{n \in \mathbb{N}}$ lies in $[\varepsilon, (1/\alpha) - \varepsilon]$, for some $\varepsilon \in]0, 1/2[$, and $F = \emptyset$. Then $\|x_n\| \rightarrow +\infty$.

Open issue 2: analysis of aperiodic networks

ASSUMPTIONS

There exist $(\omega_n)_{n \in \mathbb{N}} \in \ell_+^1$, $(\rho_n)_{n \in \mathbb{N}} \in \ell_+^1$, $(\eta_n)_{n \in \mathbb{N}} \in \ell_+^1$, and $(\nu_n)_{n \in \mathbb{N}} \in \ell_+^1$ for which the following hold, for every $i \in \{1, \dots, m\}$:

- There exists a bounded linear operator $W_i: H_{i-1} \rightarrow H_i$ such that $(\forall n \in \mathbb{N}) \|W_{i,n} - W_i\| \leq \omega_n$.
- There exists a proximal activation operator $R_i: H_i \rightarrow H_i$ such that $(\forall n \in \mathbb{N})(\forall x \in H_i) \|R_{i,n}x - R_ix\| \leq \rho_n\|x\| + \eta_n$.
- There exists $b_i \in H_i$ such that $(\forall n \in \mathbb{N}) \|b_{i,n} - b_i\| \leq \nu_n$.

$\rightsquigarrow$ Same asymptotic results

What about more general networks ?

Some references



P. L. Combettes and J.-C. Pesquet

Proximal splitting methods in signal processing

in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*,
H. H. Bauschke, R. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H.
Wolkowicz editors. Springer-Verlag, New York, pp. 185-212, 2011.



P. L. Combettes and J.-C. Pesquet

Deep Neural Network Structures Solving Variational Inequalities

<https://arxiv.org/abs/1808.07526>, 2018.



P. L. Combettes and J.-C. Pesquet

Lipschitz Certificates for Neural Network Structures Driven by Averaged Activation
Operators

<https://arxiv.org/abs/1903.01014>, 2019.