Maximizing Sparsity in NN with MIP

Research Archive

Jeff Linderoth, Changwon Lee

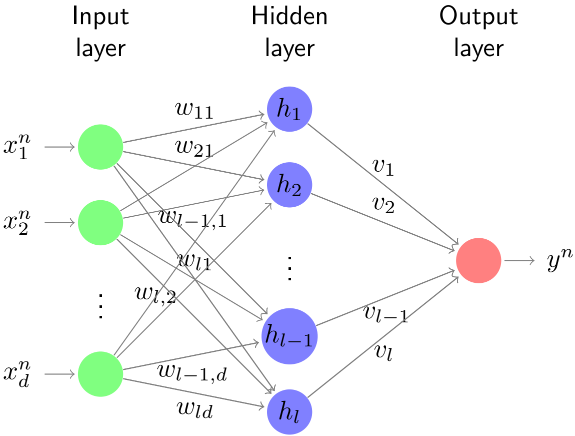
Motivated by *Rob Nowak, Fischetti and Jo*

UW-Madison, August 2025

Problem Setting

1. Given N datasets : $(\mathbf{x^1}, y^1)$,$(\mathbf{x^2}, y^2)$, $\cdots$ $(\mathbf{x^N}, y^N)$
2. Input : 'd'-dimensional
   $\mathbf{x^n} = (x_1^n, x_2^n, \ldots, x_d^n) \in \mathbb{R}^d, \quad \forall n \in [N]$
   ($\overline{\mathbf{x^n}} = (x_1^n, x_2^n, \ldots, x_d^n, 1) \in \mathbb{R}^{d+1}$)
3. w
4. $\mathbf{x^n} \rightarrow$ Affine Combination : $\mathbf{w^k}\overline{\mathbf{x^n}} \rightarrow$ Nonlinear Operator (ReLU) : $\rightarrow$ Activation of Neuron k

NN Formulation



1. $h_i^1 = max\{0, \mathbf{w} \cdot \mathbf{x^n} + b_i^1\} = max\{0, \mathbf{w} \cdot \overline{\mathbf{x^n}}\}$

MINLP Formulation 1

Decision vars:

$$\mathbf{W} \in \mathbb{R}^{l_1 \times d} : w_{ij} \in \mathbb{R}(i = 1, \ldots, l_1;\ j = 1, \ldots, d),$$

$$\mathbf{b} \in \mathbb{R}^{l_1} : b_i \in \mathbb{R}(i = 1, \ldots, l_1),$$

$$p_i^n, q_i^n \in \mathbb{R},$$

$$z_i^n \in \{0, 1\} \quad (n = 1, \ldots, N;\ i = 1, \ldots, l_1),$$

$$\mathbf{v} \in \mathbb{R}^{l_1} : v_i \in \mathbb{R}(i = 1, \ldots, l_1),$$

$$s_{ij} \in \mathbb{R} \quad (i = 1, \ldots, l_1;\ j = 1, \ldots, d),$$

$$t_{ij} \in \{0, 1\} \quad (i = 1, \ldots, l_1;\ j = 1, \ldots, d),$$

Objective:

$$\min \sum_{i=1}^{l_1} \sum_{j=1}^{d} t_{ij}$$

Constraints:

$$\begin{aligned}
\sum_{j=1}^{d} w_{ij} x_j^n + b_i &= p_i^n - q_i^n, && \forall n \in [N], i \in [l_1] \\
p_i^n, q_i^n &\geq 0, && \forall n \in [N], i \in [l_1] \\
p_i^n &\leq M(1 - z_i^n), && \forall n \in [N], i \in [l_1] \\
q_i^n &\leq M z_i^n, && \forall n \in [N], i \in [l_1] \\
\sum_{i=1}^{l_1} p_i^n v_i &= y^n, && \forall n \in [N] \\
s_{ij} &= w_{ij} v_i, && \forall i \in [l_1], j \in [d] \\
-M t_{ij} \leq s_{ij} &\leq M t_{ij}, && \forall i \in [l_1], j \in [d]
\end{aligned}$$

Normalization of Single-Layer ReLU Networks

$Lemma 1. For any given NN weights/biases$ W, b,v,
there exists an (unique) alternate weights/biases $\mathbf{W}', \mathbf{b}', \mathbf{v}'$
such that $ReLU(\mathbf{W}\mathbf{x^n} + \mathbf{b})^T\mathbf{v} = ReLU(\mathbf{W}'\mathbf{x^n} + \mathbf{b}')^T\mathbf{v}' = y^n$ for all dataset $(\mathbf{x^n}, y)$, $n \in [N]$
and $v_i' \in \{-1, 0, +1\} \quad \forall i \in [l]$.

MILP Formulation by Lemma 1

Decision vars:
$\mathbf{W} \in \mathbb{R}^{l \times d} : w_{ij} \in \mathbb{R} \quad (i = 1, \ldots, l;\ j = 1, \ldots, d),$

$\mathbf{b} \in \mathbb{R}^{l} : b_i \in \mathbb{R} \quad (i = 1, \ldots, l),$

$p_i^n, q_i^n \in \mathbb{R} \quad (n = 1, \ldots, N;\ i = 1, \ldots, l),$

$z_i^n \in \{0, 1\} \quad (n = 1, \ldots, N;\ i = 1, \ldots, l),$

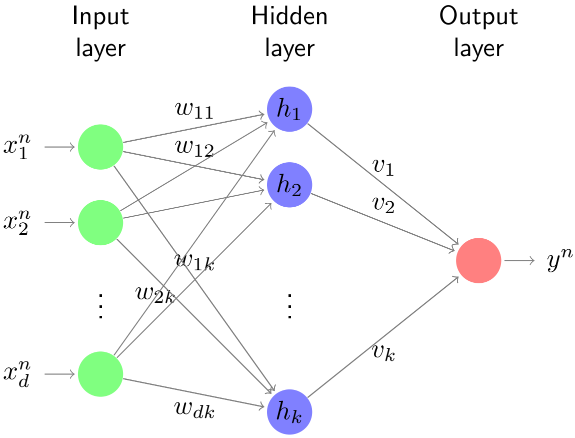
$\mathbf{v} \in \{-1, 0, +1\}^{l} : v_i^{-1}, v_i^{0}, v_i^{+1} \in \{0, 1\} \quad (i = 1, \ldots, l),$

$t_{ij} \in \{0, 1\} \quad (i = 1, \ldots, l_1;\ j = 1, \ldots, d),$

Objective:
$$\min \sum_{i=1}^{l_1} \sum_{j=1}^{d} t_{ij}$$

Constraints:
$$\begin{aligned}
\sum_{j=1}^{d} w_{ij} x_j^n + b_i &= p_i^n - q_i^n, && \forall n \in [N], i \in [l_1] \\
p_i^n, q_i^n &\geq 0, && \forall n \in [N], i \in [l_1] \\
p_i^n &\leq M(1 - z_i^n), && \forall n \in [N], i \in [l_1] \\
q_i^n &\leq M z_i^n, && \forall n \in [N], i \in [l_1] \\
\sum_{i=1}^{l} p_i^n(-v_i^{-1} + v_i^{+1}) &= y^n, && \forall n \in [N] \\
v_i^{-1} + v_i^{0} + v_i^{+1} &= 1, && \forall i \in [l] \\
-M t_{ij} \leq w_{ij} &\leq M t_{ij}, && \forall i \in [l_1], j \in [d]
\end{aligned}$$

Normalized v Formulation



1. $h_i^1 = max\{0, \mathbf{w} \cdot \mathbf{x^n} + b_i^1\} = max\{0, \mathbf{w} \cdot \overline{\mathbf{x^n}}\}$

New MILP Formulation

Given: $(\mathbf{x}^n, y^n)_{n=1}^N \in (\mathbb{R}^d \times \mathbb{R})^N : x_i^n, y^n \in \mathbb{R} \quad (n \in [N], i \in [d])$
$d, N, K \in \mathbb{N}$

Decision vars: $\theta := \{\mathbf{w_k}, b_k, v_k\}_{k=1}^K \in (\mathbb{R}^d \times \mathbb{R} \times \mathbb{R})^K$
weight 1: $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K] \in \mathbb{R}^{d \times K} : w_{ik} \in \mathbb{R} \quad (i \in [d], k \in [K])$
bias: $\mathbf{b} \in \mathbb{R}^K : b_k \in \mathbb{R} \quad (k \in [K])$
weight 2: $\mathbf{v} \in \{-1, +1\}^K : v_k \in \{-1, 1\} \ (v_k' \in \{0, 1\}, v_k = 2v_k' - 1) \quad (k \in [K])$

Output of linear combination$(\mathbf{w}_k^T \mathbf{x}^n + b_k)$(pos/neg) : $p_k^n, q_k^n \in \mathbb{R} \quad (n \in [N], k \in [K])$
Nonnegativity of Neuron(Y/N) : $z_k^n \in \{0, 1\} \quad (n \in [N], k \in [K])$
Nonzero Weight Path(Y/N) : $t_{ik} \in \{0, 1\} \quad (i \in [d], k \in [K])$

New MILP Formulation (Cont'd)

Objective(Sparsity): $$\min \sum_{i=1}^{d} \sum_{k=1}^{K} t_{ik}$$

Constraints: **Output of linear combination**

$$\sum_{i=1}^{d} w_{ik} x_i^n + b_k \; = \; p_k^n - q_k^n, \qquad \forall n \in [N], k \in [K]$$

Output of ReLu Activation

$$p_k^n, q_k^n \; \geq \; 0, \qquad \forall n \in [N], k \in [K]$$

$$p_k^n \; \leq \; M z_k^n, \qquad \forall n \in [N], k \in [K]$$

$$q_k^n \; \leq \; M(1 - z_k^n), \qquad \forall n \in [N], k \in [K]$$

Exact Data fitting

$$\sum_{k=1}^{K} p_k^n (2v_k' - 1) \; = \; y^n, \qquad \forall n \in [N]$$

Nonzero Weight Path Indicator

$$-Mt_{ik} \; \leq \; w_{ik} \; \leq \; Mt_{ik} \qquad \forall i \in [d], k \in [K]$$

McCormick Constraints

$$
\begin{aligned}
\textbf{Bilinear Constraints:} \quad & \sum_{k=1}^{K} p_k^n (2v_k' - 1) \; = \; y^n, && \forall n \in [N] \\
& \sum_{k=1}^{K} 2p_k^n v_k' - p_k^n \; = \; y^n, && \forall n \in [N] \\
\textbf{Auxiliary Variables:} \quad & r_k^n \in \mathbb{R} \;\; (r_k^n = p_k^n v_k') && \forall n \in [N], \forall k \in [K] \\
\textbf{McCormick Constraints:} \quad & (\because 0 \leq p_k^n \leq M, v_k' \in \{0,1\}) \\
& 0 \leq r_k^n \leq M v_k' && \forall n \in [N], \forall k \in [K] \\
& p_k^n - M(1 - v_k') \leq r_k^n \leq p_k^n && \forall n \in [N], \forall k \in [K]
\end{aligned}
$$

New MILP Formulation

Given: $(\mathbf{x}^n, y^n)_{n=1}^N \in (\mathbb{R}^d \times \mathbb{R})^N : x_i^n, y^n \in \mathbb{R} \quad (n \in [N], i \in [d])$
d(input dimension), N(number of data), K(width of single-hidden-layer) $\in \mathbb{N}$

Decision vars: $\theta := \{\mathbf{w_k}, b_k, v_k\}_{k=1}^K \in (\mathbb{R}^d \times \mathbb{R} \times \mathbb{R})^K$
input weight: $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K] \in \mathbb{R}^{d \times K} : w_{ik} \in \mathbb{R} \quad (i \in [d], k \in [K])$
bias: $\mathbf{b} \in \mathbb{R}^K : b_k \in \mathbb{R} \quad (k \in [K])$
output weight: $\mathbf{v} \in \{-1, +1\}^K : v_k \in \{-1, 1\} \; (v'_k \in \{0, 1\}, v_k = 2v'_k - 1) \quad (k \in [K])$

Output of linear combination$(\mathbf{w}_k^T \mathbf{x}^n + b_k)$(pos/neg) : $p_k^n, q_k^n \in \mathbb{R} \quad (n \in [N], k \in [K])$
Nonnegativity of Neuron(Y/N) : $z_k^n \in \{0, 1\} \quad (n \in [N], k \in [K])$
Nonzero input weight(Y/N) : $t_{ik} \in \{0, 1\} \quad (i \in [d], k \in [K])$
Nonzero bias(Y/N) : $s_k \in \{0, 1\} \quad (k \in [K])$

Auxiliary Variable : $r_k^n \in \mathbb{R} \; (r_k^n = p_k^n v'_k) \quad (\forall n \in [N], \forall k \in [K])$

New MILP Formulation (Cont'd)

Objective(Sparsity): $\min \sum_{i=1}^{d} \sum_{k=1}^{K} t_{ik} + \sum_{k=1}^{K} s_k$

Constraints: **Output of linear combination**

$$\sum_{i=1}^{d} w_{ik} x_i^n + b_k = p_k^n - q_k^n, \forall n \in [N], k \in [K]$$

Output of ReLU Activation

$$\begin{aligned} p_k^n, q_k^n &\geq 0, && \forall n \in [N], k \in [K] \\ p_k^n &\leq M z_k^n, && \forall n \in [N], k \in [K] \\ q_k^n &\leq M(1 - z_k^n), && \forall n \in [N], k \in [K] \end{aligned}$$

Exact Data fitting

$$\sum_{k=1}^{K} 2r_k^n - p_k^n = y^n, \qquad \forall n \in [N]$$

McCormick Constraints

$$\begin{aligned} &0 \leq r_k^n \leq M v_k' && \forall n \in [N], \forall k \in [K] \\ &p_k^n - M(1 - v_k') \leq r_k^n \leq p_k^n && \forall n \in [N], \forall k \in [K] \end{aligned}$$

Nonzero Weight/Bias Indicator

$$\begin{aligned} -Mt_{ik} \leq w_{ik} \leq Mt_{ik} && \forall i \in [d], k \in [K] \\ -Ms_k \leq b_k \leq Ms_k && \forall k \in [K] \end{aligned}$$

MILP Formulation (Cont'd

Linear Objective $\min \sum_{i=1}^{d} \sum_{k=1}^{K} t_{ik} + \sum_{k=1}^{K} s_k$

Total number of variables:

Symmetry Handling

Permutations: $\min \sum_{i=1}^{d} \sum_{k=1}^{K} t_{ik}$

References I