Maximizing Sparsity in NN with MIP
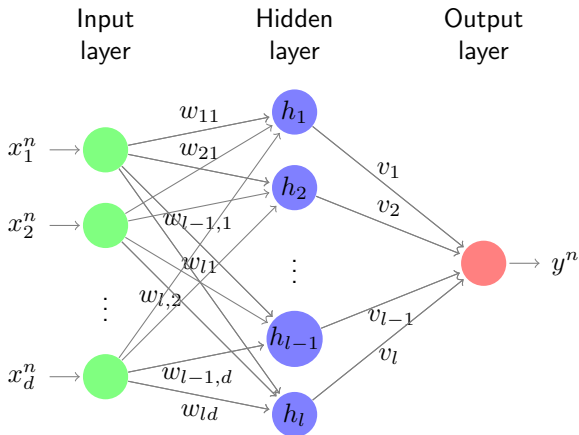## Research Archive

Jeff Linderoth, Changwon Lee

Motivated by *Rob Nowak, Fischetti and Jo*
UW-Madison, August 2025

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

Input layer   Hidden layer   Output layer

x_1^n, x_2^n, x_d^n

w_{11}, w_{21}, $w_{l-1,1}$, w_{l1}, $w_{l,2}$, $w_{l-1,d}$, w_{ld}

h_1, h_2, h_{l-1}, h_l

v_1, v_2, v_{l-1}, v_l

y^n

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

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: $\min \sum\limits_{i=1}^{l_1} \sum\limits_{j=1}^{d} t_{ij}$

Constraints:

$$\sum_{j=1}^{d} w_{ij} x_j^n + b_i = p_i^n - q_i^n, \qquad \forall n \in [N], i \in [l_1]$$

$$p_i^n, q_i^n \geq 0, \qquad \forall n \in [N], i \in [l_1]$$

$$p_i^n \leq M(1 - z_i^n), \qquad \forall n \in [N], i \in [l_1]$$

$$q_i^n \leq M z_i^n, \qquad \forall n \in [N], i \in [l_1]$$

$$\sum_{i=1}^{l_1} p_i^n v_i = y^n, \qquad \forall n \in [N]$$

$$s_{ij} = w_{ij} v_i, \qquad \forall i \in [l_1], j \in [d]$$

$$-M t_{ij} \leq s_{ij} \leq M t_{ij}, \qquad \forall i \in [l_1], j \in [d]$$

$Lemma 1. For any given NN weights/biases$ W, b,v,
there exists an (unique) alternate weights/biases $\mathbf{W}', \mathbf{b}', \mathbf{v}'$
such that $ReLU(\mathbf{Wx^n} + \mathbf{b})^T \mathbf{v} = ReLU(\mathbf{W'x^n} + \mathbf{b'})^T \mathbf{v'} = y^n$ 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:

$$\sum_{j=1}^{d} w_{ij} x_j^n + b_i = p_i^n - q_i^n, \qquad \qquad \forall n \in [N], i \in [l_1]$$

$$p_i^n, q_i^n \geq 0, \qquad \qquad \forall n \in [N], i \in [l_1]$$

$$p_i^n \leq M(1 - z_i^n), \qquad \qquad \forall n \in [N], i \in [l_1]$$
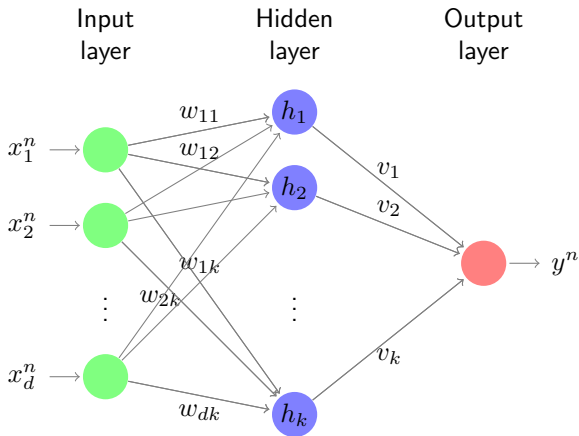
$$q_i^n \leq M z_i^n, \qquad \qquad \forall n \in [N], i \in [l_1]$$

$$\sum_{i=1}^{l} p_i^n(-v_i^{-1} + v_i^{+1}) = y^n, \qquad \qquad \forall n \in [N]$$

$$v_i^{-1} + v_i^0 + v_i^{+1} = 1, \qquad \qquad \forall i \in [l]$$

$$-M t_{ij} \leq w_{ij} \leq M t_{ij}, \qquad \qquad \forall i \in [l_1], j \in [d]$$

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

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) $: t_{ik} \in \{0, 1\} \quad (i \in [d], k \in [K])$

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

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

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

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

Exact Data fitting

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

Nonzero Weight Path Indicator

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

McCormick Constraints

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

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

Auxiliary Variables: $r_k^n \in \mathbb{R} \ (r_k^n = p_k^n v_k') \qquad \forall n \in [N], \forall k \in [K]$

McCormick Constraints: $(\because 0 \leq p_k^n \leq M, v_k' \in \{0, 1\})$

$0 \leq r_k^n \leq M v_k' \qquad \forall n \in [N], \forall k \in [K]$

$p_k^n - M(1 - v_k') \leq r_k^n \leq p_k^n \quad \forall n \in [N], \forall k \in [K]$

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) $: t_{ik} \in \{0, 1\} \quad (i \in [d], k \in [K])$

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

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

Constraints: **Output of linear combination**

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

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

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

McCormick Constraints

$$0 \leq r_k^n \leq M v_k' \qquad \forall n \in [N], \forall k \in [K]$$

$$p_k^n - M(1 - v_k') \leq r_k^n \leq p_k^n \quad \forall n \in [N], \forall k \in [K]$$

Nonzero Weight Path Indicator

$$-M t_{ik} \leq w_{ik} \leq M t_{ik} \quad \forall i \in [d], k \in [K]$$

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