On the Inductive Bias of Neural Networks for Learning Read-once DNFs (Supplementary material)

Ido Bronstein[1] **Alon Brutzkus**[1] **Amir Globerson**[1]

[1]Blavatnik School of Computer Science , Tel Aviv University, Israel

Here we provide complete proofs for the results in the main paper, as well as additional empirical results.

1 GRADIENT FLOW - ASSUMPTIONS

When gradient flow is implemented on non-differentiable functions (e.g., ReLU) the implementation can choose from among a set of possible sub-differentials. Here we define which of these our analysis will use. This choice corresponds to the common way gradient methods are implemented in practice for the ReLU function.

Recall that the gradient flow step is $\frac{d\boldsymbol{\theta}^{(t)}}{dt} \in -\partial^\circ L(\boldsymbol{\theta}^{(t)})$ for a.e. t, where:

$$\partial^\circ f(\boldsymbol{x}) = \text{conv}\left\{\lim_{k\to\infty} \nabla f(\boldsymbol{x}_k) \mid \boldsymbol{x}_k \to \boldsymbol{x} \text{ and } f \text{ is differentiable at } \boldsymbol{x}_k\right\} \tag{7}$$

is the Clarke's subdifferential.

As discussed in the main text, we will assume that the gradient flow step selects a specific vector in the subdifferential. This is done by setting the subgradient of ReLU at 0 in advance to a constant value $a \in [0,1]$. Namely, this value of the subgradient is used for all neurons and in all iterations. Usually a is set to be either 0 or 1.

Formally, for each $i \in r$ we denote $\frac{d\boldsymbol{w}_i^{(t)}}{dt} = \frac{1}{m}\sum_{\boldsymbol{x}\in\mathbb{S}} \frac{d\boldsymbol{w}_i^{(t)}(\boldsymbol{x})}{dt}$. Here, $\frac{d\boldsymbol{w}_i^{(t)}(\boldsymbol{x})}{dt}$ is the gradient update of $\boldsymbol{w}_i^{(t)}$ restricted to the summand that depends on $\boldsymbol{x}$ in L (Eq. (2) of the main paper). Similarly, we denote $\frac{db_i^{(t)}}{dt} = \frac{1}{m}\sum_{\boldsymbol{x}\in\mathbb{S}} \frac{db_i^{(t)}(\boldsymbol{x})}{dt}$. For our result, we need the following technical assumption.

Assumption 1.1. *There exists an $a \in [0,1]$ such that for every step $t > 0$, every neuron $i \in [r]$, and every sample $\boldsymbol{x} \in \mathbb{S}$ if $\boldsymbol{w}_i^{(t)} \cdot \boldsymbol{x} + b_i^{(t)} = 0$ then $\frac{d\boldsymbol{w}_i^{(t)}(\boldsymbol{x})}{dt} = ay\ell'\left(yN(\boldsymbol{x};\boldsymbol{\theta}^{(t)})\right)\boldsymbol{x}$ and $\frac{db_i^{(t)}}{dt} = ay\ell'\left(yN(\boldsymbol{x};\boldsymbol{\theta}^{(t)})\right)$.*

2 PROOF OF THEOREM 5.1

We first need the following notation and recall the KKT conditions in our context. Let $\partial^\circ\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}_l + b) \subseteq \mathbb{R}^{D+1}$ be the subdifferential of neuron i given input $\boldsymbol{x}_l$. It holds that:

$$\partial^\circ\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}_l + b) = \begin{cases} \{(\boldsymbol{x}_l, 1)\} & \text{if } \boldsymbol{w}_i \cdot \boldsymbol{x}_l + b_i > 0 \\ \{(\boldsymbol{0}, 0)\} & \text{if } \boldsymbol{w}_i \cdot \boldsymbol{x}_l + b_i < 0 \\ [0,1]^{D+1} & \text{if } \boldsymbol{w}_i \cdot \boldsymbol{x}_l + b_i = 0 \end{cases}$$

KKT conditions: A feasible point $\boldsymbol{\theta} = (\boldsymbol{W}, \boldsymbol{b}, c)$ of the min norm problem (Eq. (4) of the main paper) is a KKT point if there exist $\lambda_1, \ldots \lambda_m \geq 0$ such that:

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

1. **Stationarity**:

$$\forall i \in [r],\ \ \boldsymbol{w}_i = \sum_{l\in[m]} \lambda_l y_l \boldsymbol{h}_{il} \text{ and } b_i = \sum_{l\in[m]} \lambda_l y_l g_{il} \tag{8}$$

and

$$c = \sum_{j\in[m]} \lambda_j y_j \tag{9}$$

where $(\boldsymbol{h}_{il}, g_{il}) \in \partial^\circ \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}_l + b)$.

2. **Complementary slackness**: if $y_i N(\boldsymbol{x}_i, \theta) > 1$, then $\lambda_i = 0$.

We now proceed to prove the theorem. In the first part, we show that by Assumption 1.1, we can restrict the possible values of $\boldsymbol{h}_{il}, g_{il}$ for a KKT point that GF converges to. In the second part, we prove properties of neurons that memorize samples. In the third part, we use the previous parts to show that memorizing solutions cannot be KKT points.

Part 1: In this part we prove the following lemma.

Lemma 2.1. *Assume that Assumption 1.1 holds and GF converges to a KKT point with parameters $\boldsymbol{h}_{il}$ and g_{il} for $1 \le i \le r$ and $1 \le l \le n$, as defined in Eq. (8). Then, $(\boldsymbol{h}_{il}, g_{il}) \in \{(\boldsymbol{x}_l, 1), (a\boldsymbol{x}_l, a), (0, 0)\}$.*

Proof. For each $1 \le l \le n$, $1 \le i \le r$ and $t > 0$ let $\left(\boldsymbol{h}_{il}^{(t)}, g_{il}^{(t)}\right) \in \partial^\circ \sigma\left(\boldsymbol{w}_i^{(t)} \cdot \boldsymbol{x}_l + b_i^{(t)}\right)$ be the corresponding values of the GF step at time t in the subdifferential $\partial^\circ \sigma\left(\boldsymbol{w}_i^{(t)} \cdot \boldsymbol{x}_l + b_i^{(t)}\right)$. By inspecting the proofs of Lyu & Li (2020) and Dutta et al. (2013), we see that $(\boldsymbol{h}_{il}, g_{il})$ is equal to the limit of a convergent subsequence of $\left\{\left(h_{il}^{(t_j)}, g_{il}^{(t_j)}\right)\right\}_{j=0}^{\infty}$.

By assumption 1.1, we know that for each j, $\left(\boldsymbol{h}_{il}^{(t_j)}, g_{il}^{(t_j)}\right) \in \{(\boldsymbol{x}_l, 1), (a\boldsymbol{x}_l, a), (\boldsymbol{0}, 0)\}$. Therefore, the limit also satisfies $(\boldsymbol{h}_{il}, g_{il}) \in \{(\boldsymbol{x}_l, 1), (a\boldsymbol{x}_l, a), (\boldsymbol{0}, 0)\}$.

□

Part 2: We will first need the following definition.

Definition 2.1. *Given a sample $\hat{\boldsymbol{x}} \in \mathcal{X}$ and index $j \in [D]$, the sample with Hamming distance one from $\hat{\boldsymbol{x}}$ at index j is defined as $\mathcal{H}(\hat{\boldsymbol{x}}, j) \in \mathcal{X}$ and satisfies the following:*

$$\mathcal{H}(\hat{\boldsymbol{x}}, j)_j = -x_j \text{ and } \forall j' \in [D]\backslash\{j\}\ \ \mathcal{H}(\hat{\boldsymbol{x}}, j)_{j'} = x_{j'} \tag{10}$$

The set of all samples with Hamming distance one from $\hat{\boldsymbol{x}}$ is defined as: $\Psi(\hat{\boldsymbol{x}}) = \{\boldsymbol{x}' \in \mathcal{X} \mid \exists j \in [D]\ \ \mathcal{H}(\hat{\boldsymbol{x}}, j) = \boldsymbol{x}'\}$. Note, $|\Psi(\boldsymbol{x})| = D$

Using this definition, we rephrase Lemma 5.1 of the main text and prove those properties of memorizing neurons (Definition 5.1 of the main paper):

Lemma 2.2. *Let $D > 2$. If a neuron $i \in [r]$ memorizes a sample $\hat{\boldsymbol{x}} \in \mathbb{S}_x$, then it satisfies the following properties:*

1. $\hat{x}_j = \text{sign}(w_{ij})$ *for all* $1 \le j \le D$.
2. *For* $\boldsymbol{x} \in \mathcal{X}$ *if* $\boldsymbol{w}_i \cdot \boldsymbol{x} + b_i = 0$ *then* $\boldsymbol{x} \in \Psi(\hat{\boldsymbol{x}})$.
3. $b_i < 0$

Proof. **Property 1:** Assume by contradiction that there exists $j \in [D]$ such that $\text{sign}(w_{ij}) \neq \hat{x}_j$. Then $\boldsymbol{w}_i \cdot \mathcal{H}(\hat{\boldsymbol{x}}, j) + b_i \ge \boldsymbol{w}_i \cdot \hat{\boldsymbol{x}} + b_i > 0$, in contradiction to the memorization assumption in Eq. (5) of the main paper.

Property 2: Assume by contradiction that there exists $\boldsymbol{x} \in \mathcal{X}\backslash(\Psi(\hat{\boldsymbol{x}}) \cup \{\hat{\boldsymbol{x}}\})$ such that $\boldsymbol{w}_i \cdot \boldsymbol{x} + b_i = 0$. We define $J = \{j \in [D] \mid x_j = -\hat{x}_j\}$ and $j' \in J$ for some index in J. By Property 1, the sample $\widetilde{\boldsymbol{x}} = \mathcal{H}(\hat{\boldsymbol{x}}, j')$ satisfies the

following:

$$\begin{aligned}\boldsymbol{w}_i \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \widetilde{\boldsymbol{x}} &= \sum_{j\in[D]\setminus J} w_{ij}x_j + \sum_{j\in J} w_{ij}x_j - \sum_{j\in[D]\setminus\{j^\theta\}} w_{ij}\widetilde{x_j} - w_{ij^\theta}\widetilde{x_j} \\ &= \sum_{j\in[D]\setminus J} w_{ij}\hat{x}_j - \sum_{j\in J} w_{ij}\hat{x}_j - \sum_{j\in[D]\setminus\{j^\theta\}} w_{ij}\hat{x}_j + w_{ij^\theta}\hat{x}_j \\ &= \sum_{j\in[D]\setminus J} |w_{ij}| - \sum_{j\in J}|w_{ij}| - \sum_{j\in[D]\setminus\{j^\theta\}} |w_{ij}| + |w_{ij^\theta}| = -2\sum_{j\in J\setminus\{j^\theta\}} |w_{ij}|\end{aligned} \tag{11}$$

Since $\boldsymbol{x} \notin \Psi(\hat{\boldsymbol{x}})$, it holds that $J\setminus\{j'\} \neq \emptyset$. Furthermore, by Property 1, $\forall j \in J\setminus\{j'\}\ w_{ij} \neq 0$. Thus, $\sum_{j\in J\setminus\{j^\theta\}} |w_{ij}| > 0$ which implies that $\boldsymbol{w}_i \cdot \boldsymbol{x} < \boldsymbol{w}_i \cdot \widetilde{\boldsymbol{x}}$. We know that $\boldsymbol{w}_i \cdot \boldsymbol{x} + b_i = 0$, therefore $\boldsymbol{w}_i \cdot \widetilde{\boldsymbol{x}} + b_i > 0$. This contradicts the memorization assumption in Eq. (5) of the main paper.

Property 3: Assume by contradiction that $b_i \geq 0$. We define $j' = \arg\min_{j\in[D]}\{|w_{ij}|\}$. For the sample $\widetilde{\boldsymbol{x}} = \mathcal{H}(\hat{\boldsymbol{x}}, j')$, the following holds by Property 1:

$$\boldsymbol{w}_i \cdot \widetilde{\boldsymbol{x}} = \sum_{j\in[D]\setminus\{j^\theta\}} w_{ij}\hat{x}_j - w_{ij^\theta}\hat{x}_{j^\theta} = \sum_{j\in[D]\setminus\{j^\theta\}} |w_{ij}| - |w_{ij^\theta}| > 0 \tag{12}$$

Since $D > 2$, we have $\sum_{j\in[D]\setminus\{j^\theta\}} |w_{ij}| - |w_{ij^\theta}| > 0$. Thus, $\boldsymbol{w}_i \cdot \widetilde{\boldsymbol{x}} + b_i > 0$ which contradicts the memorization assumption in Eq. (5) of the main paper. □

Part 3: Now, we can proceed to prove Theorem 5.1.

Consider a network with parameters $\bar{\boldsymbol{\theta}} = (\bar{\boldsymbol{W}}, \bar{\boldsymbol{b}}, \bar{c})$, neuron $i \in [r]$ and a sample $(\hat{\boldsymbol{x}}, \hat{y}) \in \mathbb{S}$ such that Eq. (5) of the main paper holds (i.e., the neuron i memorizes the sample $\hat{x}$). We assume by contradiction that there exists an initialization $\boldsymbol{\theta}^{(0)}$ such that if we run gradient flow from $\boldsymbol{\theta}^{(0)}$ using μ then the weights $\boldsymbol{\theta}^{(t)}$ will converge to $\bar{\boldsymbol{\theta}}$. According to the results of Lyu & Li (2020); Ji & Telgarsky (2020), we know that there exists $\alpha > 0$ such that $\boldsymbol{\theta} = (\boldsymbol{W}, \boldsymbol{b}, c) = \alpha\bar{\boldsymbol{\theta}}$ is a KKT point of Eq. (4) of the main paper and Eq. (8) holds for $\boldsymbol{\theta}$. Note that for $\boldsymbol{\theta}$, neuron i memorizes the sample $(\hat{\boldsymbol{x}}, \hat{y})$ as well.

Given a sample $\boldsymbol{x}_l \in \mathbb{S}_x$ we can see that the following holds:

1. If $\boldsymbol{x}_l = \hat{\boldsymbol{x}}$ then $(\boldsymbol{h}_{il}, g_{il}) = \{(\hat{\boldsymbol{x}}, 1)\}$.
2. If $\boldsymbol{x}_l \in \Psi(\hat{\boldsymbol{x}})$ then $(\boldsymbol{h}_{il}, g_{il}) = \{(a\boldsymbol{x}_l, a)\}$ where $a \in [0, 1]$ by Assumption 1.1.
3. If $\boldsymbol{x}_l \notin \Psi(\hat{\boldsymbol{x}})$ then $(\boldsymbol{h}_{il}, g_{il}) = \{(\boldsymbol{0}, 0)\}$ by Property 2 of Lemma 2.2.

We will show that for every $\lambda_1, \ldots, \lambda_{|\mathbb{S}|} \geq 0$, Eq. (8) does not hold. We can assume without loss of generality that all the samples in $\boldsymbol{x}_l \in \Psi(\boldsymbol{x}) \cap \mathbb{S}_x$ are support vectors and $(\boldsymbol{h}_{il}, g_{il}) = (\boldsymbol{x}_l, 1)$. This is because if one of the samples is not a support vector then we can take $\lambda_l = 0$. Furthermore, if $(\boldsymbol{h}_{il}, g_{il}) = (\boldsymbol{0}, 0)$ then we can take $\lambda_l = 0$ and if $(\boldsymbol{h}_i, g_i) = (a\boldsymbol{x}_l, a)$ for $a > 0$ we can set $\frac{\lambda_l}{a}$ instead of λ_l. Under this assumption we can write Eq. (8) for $\tilde{\boldsymbol{\theta}}$ using only $\lambda_1, \ldots, \lambda_D$ and $\hat{\lambda}$ that correspond to the samples of $\Psi(\hat{\boldsymbol{x}})$ and $\hat{\boldsymbol{x}}$, respectively:

$$\boldsymbol{w}_i = \sum_{\boldsymbol{x}_l\in\Psi(\hat{\boldsymbol{x}})\cap\mathbb{S}_x} \lambda_l y_l \boldsymbol{x}_l + \hat{\lambda}\hat{y}\hat{\boldsymbol{x}} \quad \text{and} \quad b_i = \sum_{\boldsymbol{x}_l\in\Psi(\hat{\boldsymbol{x}})\cap\mathbb{S}_x} \lambda_l y_l + \hat{\lambda}\hat{y} \tag{13}$$

Assume by contradiction that there exists $\boldsymbol{x}_{l^\theta} = \mathcal{H}(\hat{\boldsymbol{x}}, j')$ such that $\boldsymbol{x}_{l^\theta} = \Psi(\hat{\boldsymbol{x}})\setminus\mathbb{S}_x$. The following holds by Eq. (13) and Definition 2.1:

$$\begin{aligned}&(1)\ \ w_{ij^\theta} = \hat{\lambda}\hat{y}\hat{x}_{j^\theta} + \sum_{\boldsymbol{x}_l\in\Psi(\hat{\boldsymbol{x}})\cap\mathbb{S}_x} \lambda_l y_l \hat{x}_{j^\theta} \\ &(2)\ \ b_i = \hat{\lambda}\hat{y} + \sum_{\boldsymbol{x}_l\in\Psi(\hat{\boldsymbol{x}})\cap\mathbb{S}_x} \lambda_l y_l\end{aligned} \tag{14}$$

Using the first property in Lemma 2.2:

$$(1)\ \ w_{ij^\vartheta} = \hat{\lambda}\hat{y}\,\mathrm{sign}(w_{ij^\vartheta}) + \sum_{\boldsymbol{x}_l \in \Psi(\hat{\boldsymbol{x}}) \cap \mathbb{S}_x} \lambda_l y_l\,\mathrm{sign}(w_{ij^\vartheta})$$
$$(2)\ \ b_i = \hat{\lambda}\hat{y} + \sum_{\boldsymbol{x}_l \in \Psi(\hat{\boldsymbol{x}}) \cap \mathbb{S}_x} \lambda_l y_l \tag{15}$$

Therefore,

$$(1)\ \ |w_{ij^\vartheta}| = \hat{\lambda}\hat{y} + \sum_{\boldsymbol{x}_l \in \Psi(\hat{\boldsymbol{x}}) \cap \mathbb{S}_x} \lambda_l y_l$$
$$(2)\ \ b_i = \hat{\lambda}\hat{y} + \sum_{\boldsymbol{x}_l \in \Psi(\hat{\boldsymbol{x}}) \cap \mathbb{S}_x} \lambda_l y_l \tag{16}$$

This means that $0 \leq |w_{ij^\vartheta}| = b_i$, which is in contradiction to the third property of Lemma 2.2. Therefore, we can assume from now on that $\Psi(\hat{\boldsymbol{x}}) \subseteq \mathbb{S}_x$, and we can write the KKT conditions as follows:

$$\boldsymbol{w}_i = \sum_{\boldsymbol{x}_l \in \Psi(\hat{\boldsymbol{x}})} \lambda_l y_l \boldsymbol{x}_l + \hat{\lambda}\hat{y}\hat{\boldsymbol{x}} \ \text{ and } \ b_i = \sum_{\boldsymbol{x}_l \in \Psi(\hat{\boldsymbol{x}})} \lambda_l y_l + \hat{\lambda}\hat{y} \tag{17}$$

Given $\boldsymbol{x}_{l^\vartheta} = \mathcal{H}(\hat{\boldsymbol{x}}, j')$, the following holds by Eq. (17) and Definition 2.1:

$$(1)\ \ w_{ij^\vartheta} = \hat{\lambda}\hat{y}\hat{x}_{j^\vartheta} + \sum_{\boldsymbol{x}_l \in \Psi(\hat{\boldsymbol{x}}) \setminus \{\boldsymbol{x}_{l^\vartheta}\}} \lambda_l y_l \hat{x}_{j^\vartheta} - \lambda_{l_1} y_{l_1} \hat{x}_{j^\vartheta}$$
$$(2)\ \ b_i = \hat{\lambda}\hat{y} + \sum_{\boldsymbol{x}_l \in \Psi(\hat{\boldsymbol{x}})} \lambda_l y_l \tag{18}$$

Using the first property in Lemma 2.2:

$$(1)\ \ w_{ij^\vartheta} = \hat{\lambda}\hat{y}\,\mathrm{sign}(w_{ij^\vartheta}) + \sum_{\boldsymbol{x}_l \in \Psi(\hat{\boldsymbol{x}}) \setminus \{\boldsymbol{x}_{l^\vartheta}\}} \lambda_l y_l\,\mathrm{sign}(w_{ij^\vartheta}) - \lambda_{l_1} y_{l^\vartheta}\,\mathrm{sign}(w_{ij^\vartheta})$$
$$(2)\ \ b_i = \hat{\lambda}\hat{y} + \sum_{\boldsymbol{x}_l \in \Psi(\hat{\boldsymbol{x}})} \lambda_l y_l \tag{19}$$

Therefore,

$$(1)\ \ |w_{ij^\vartheta}| = \hat{\lambda}\hat{y} + \sum_{\boldsymbol{x}_l \in \Psi(\hat{\boldsymbol{x}}) \setminus \{\boldsymbol{x}_{l^\vartheta}\}} \lambda_l y_l - \lambda_{l^\vartheta} y_{l^\vartheta}$$
$$(2)\ \ b_i = \hat{\lambda}\hat{y} + \sum_{\boldsymbol{x}_l \in \Psi(\hat{\boldsymbol{x}})} \lambda_l y_l \tag{20}$$

The result of subtracting $(2) - (1)$ is:

$$b_i - |w_{ij^\vartheta}| = 2\lambda_{l^\vartheta} y_{l^\vartheta} \tag{21}$$

Next we show that it must hold that $y_{l^\vartheta} = -1$. To see this, note that the third property of Lemma 2.2 implies $b_i < 0$ and therefore $b_i - |w_{ij^\vartheta}| < 0$. Assuming in contradiction that $y_{l^\vartheta} = 1$, the RHS of Eq. (21) satisfies $0 \leq 2\lambda_{l^\vartheta} y_{l^\vartheta}$. But we just saw that the LHS satisfies $b_i - |w_{ij^\vartheta}| < 0$. We therefore have a contradiction and conclude that $y_{l^\vartheta} = -1$. We can conclude that $\Psi(\hat{\boldsymbol{x}})$ contains only negative samples.

Next we argue that $\hat{y} = -1$. To see this, assume in contradiction that $\hat{y} = 1$. Then there exists a $n \in [K]$ such that $\hat{\boldsymbol{x}}$ satisfies the term t_n^*. Due to the fact that $K \geq 2$, we know that there exists $j \in [D] \setminus \mathbb{A}_n$. Then, $\mathcal{H}(\hat{\boldsymbol{x}}, j) \in \Psi(\hat{\boldsymbol{x}})$ is a positive sample in contradiction to the fact that $\Psi(\hat{\boldsymbol{x}})$ contains only negative samples, and therefore $\hat{y} = -1$

By Eq. (21) we know that for every $j \in [D]$ a sample $\boldsymbol{x}_l = \mathcal{H}(\hat{\boldsymbol{x}}, j)$ satisfies the following: $\lambda_l = \frac{1}{2}(|w_{ij}| - b_i)$. If we assign this in Eq. (20), we get:

$$|w_{ij^\vartheta}| = \hat{\lambda}\hat{y} + \sum_{\boldsymbol{x}_l \in \Psi(\hat{\boldsymbol{x}}) \setminus \{\boldsymbol{x}_{l^\vartheta}\}} \frac{1}{2}(b_i - |w_{ij}|) - \frac{1}{2}(b_i - |w_{ij^\vartheta}|) \tag{22}$$

Therefore,

$$0 = \hat{y} + \frac{1}{2}(D - 2)b_i - \frac{1}{2}\|w_i\|_1 \tag{23}$$

But this results in a contradiction, because we know that $\hat{y} \geq 0$, $\frac{1}{2}(D-2)b_i < 0$ and $\frac{1}{2}\|w_i\|_1 < 0$.

Thus, we conclude that gradient flow cannot converge to memorizing solutions.

3 PROOF OF THEOREM 6.1

We prove the theorem in several parts. We first prove properties of a perfect solution (Section 3.1). In Section 3.2 we prove several results regarding the bias threshold. In Section 3.3 we prove auxiliary lemmas and in Section 3.4 we prove the alignment of the neurons of the optimal solutions to the terms of the DNF. We conclude the proof in Section 3.5.

3.1 A SIMPLE PROPERTY OF PERFECT SOLUTIONS

Recall the definitions $S_+ = \{x \mid (x,1) \in S\}$ and $S_- = \{x \mid (x,-1) \in S\}$. We first need the following definitions. We say that a solution (W,b) satisfies the MIN_+ property if for any positive point $x \in S_+$ there exists $I \subseteq [r]$ such that $\sum_{i \in I} w_i \cdot x + b_i \geq 2$. We say that a solution satisfies the MIN_- property if for any negative point $x \in S_-$ and for all $i \in [r]$, $w_i \cdot x + b_i \geq 0$.

Lemma 3.1. (W,b) is a perfect solution if and only if (W,b) satisfies MIN_+ and MIN_-.

Proof. If $\theta = (W,b)$ is a perfect solution, then for all $(x,y) \in S$, $y \cdot N(x;W,b) \geq 1$. Therefore, if $y = 1$, $\sum_{i \in [r]} (w_i \cdot x + b_i) \geq 2$. Thus, there exists $I \subseteq [r]$ such that $\sum_{i \in I} w_i \cdot x + b_i \geq 2$ and the MIN_+ holds. If $y = -1$, then $\sum_{i \in [r]} (w_i \cdot x + b_i) \leq 0$ and therefore for all $i \in [r]$, $w_i \cdot x + b_i \leq 0$. The other direction follows similarly. □

We note that one direct consequence of Lemma 3.1 is that given a negative $x \in S_-$ is activated by a neuron $i \in [r]$, i.e., $w_i \cdot x + b_i > 0$, then the MIN_- property doesn't hold, and is not a perfect solution.

3.2 PROOF OF LEMMA 6.1

In this section we show that when $S_X = X$, the bias of any neuron in a perfect solution is upper bounded by a certain value which we call the *bias threshold*. To simplify the formulation of this section we define the following:

Definition 3.1. We define the set of indices which are not active in any term as the noisy indices and denote them by $A_{K+1} = [D] \setminus \bigcup_{n \in [K]} A_n$.

Definition 3.2. For each term $n \in [K]$ of θ and $i \in [r]$ define $V_n(w_i) = \max\left(\min_{j \in A_n} \{w_{ij}\}, 0\right)$.

Definition 3.3. The bias threshold for a weight w is $BT(w) = -\|w\|_1 + 2\sum_{n \in [K]} V_n(w)$.

Note, $BT(w) \leq 0$ because every term includes at least 2 literals. Using those definitions we rephrase Lemma 6.1 and prove it:

Lemma 3.2. Assume that $S_X = X$. (W,b) satisfies that $\forall i \in [r]$ $b_i \leq BT(w_i)$ and satisfies the MIN_+ property if and only if the network is a perfect solution.

Proof. We will show that given a neuron (w,b), there is a negative sample $x \in X$ for which $w \cdot x + b > 0$ if and only if $b > BT(w)$. By showing that and using the assumption $S_X = X$, we can conclude that $\forall i \in [r]$ $b_i \leq BT(w_i)$ if and only if the MIN_- property holds. By Combining this with Lemma 3.1, we can prove our claim.

Given neuron $(w; b)$, we define the minimum index of a term $n \in [K]$ as $J_n = \arg\min_{j \in A_n} \{w_j\}$. Consider a sample $\hat{x} \in X$ that is defined by:

$$\hat{x}_j = \begin{cases} \text{sign}(w_j) & \exists n \in [K] : V_n(w) > 0 \wedge j = J_n \\ \text{sign}(w_j) & \text{otherwise} \end{cases} \tag{24}$$

For every term $n \in [K]$, if $V_n(w) > 0$ then $\hat{x}_{J_n} = \text{sign}(w_{J_n}) = 1$. Otherwise, $V_n(w) = 0$ and there exists $j \in A_n$ such that $w_j < 0$, i.e., $\hat{x}_j = \text{sign}(w_j) = -1$. In any case $\hat{x}^\top t_n < |A_n|$. Therefore, the label of this sample is negative and denote it by $\hat{y} = -1$.

We show that $w^\top \hat{x} = BT(w)$ by,

$$\begin{aligned}
w^\top \hat{x} &= \sum_{j \in [D]} w_j \hat{x}_j = \sum_{n \in [K+1]} \Big[\sum_{j \in A_n} w_j \hat{x}_j \Big] \\
&= \sum_{n \in [K] \text{ and } V_n(w) > 0} \; \sum_{j \in A_n \setminus \{J_n\}} w_j \, \text{sign}(w_j) \; - \; w_{J_n} \, \text{sign}(w_{J_n}) \\
&\quad + \sum_{n \in [K] \text{ and } V_n(w) = 0} \; \sum_{j \in A_n} w_j \, \text{sign}(w_j) + \sum_{j \in A_{K+1}} w_j \, \text{sign}(w_j) \\
&= \sum_{n \in [K] \text{ and } V_n(w) > 0} \; \sum_{j \in A_n \setminus \{J_n\}} |w_j| \; - \; |w_{J_n}| + \sum_{n \in [K] \text{ and } V_n(w) = 0} \; \sum_{j \in A_n} |w_j| + \sum_{j \in A_{K+1}} |w_j| \\
&= \sum_{n \in [K] \text{ and } V_n(w) > 0} \; \sum_{j \in A_n} |w_j| \; - \; 2|w_{J_n}| + \sum_{n \in [K] \text{ and } V_n(w) = 0} \; \sum_{j \in A_n} |w_j| + \sum_{j \in A_{K+1}} |w_j| \\
&= \sum_{n \in [K]} \; \sum_{j \in A_n} |w_j| \; - \; 2|w_{J_n}| \, V_n(w) + \sum_{j \in A_{K+1}} |w_j| = \sum_{j \in [D]} |w_j| \; - \; 2 \sum_{n \in [K]} V_n(w) \\
&= \|w\|_1 \; - \; 2 \sum_{n \in [K]} V_n(w) = BT(w)
\end{aligned} \tag{25}$$

For the first direction, if $b > BT(w)$ then $w^\top \hat{x} + b = BT(w) + b > 0$, as desired.

In the second direction, assume that there is a negative sample $x \in X$ such that $w^\top x + b > 0$. We will show that $x^\top w \geq \hat{x}^\top w$. Every term $n \in [K]$ satisfies for all $j \in A_n \setminus \{J_n\}$:

$$x_j w_j \; - \; |w_j| = \text{sign}(w_j) w_j = \hat{x}_j w_j \tag{26}$$

If $V_n(w) = 0$, then the index $j = J_n$ also satisfies Eq. (26), by the definition of $\hat{x}$. Otherwise, $V_n(w) > 0$ and we know that there exists $j' \in A_n$ such that $x_{j'} = -1$ (since x is negative), and $w_{J_n}, w_{j'} > 0$. If $J_n = j'$, then $x_{j'} w_{j'} = \hat{x}_{J_n} w_{J_n}$. Otherwise, the following holds:

$$x_{j'} w_{j'} + x_{J_n} w_{J_n} \qquad w_{j'} + w_{J_n} \qquad 0 \qquad w_{j'} \qquad w_{J_n} \qquad \hat{x}_{j'} w_{j'} + \hat{x}_{J_n} w_{J_n} \tag{27}$$

Note that every index in $j \in A_{K+1}$ satisfies Eq. (26) as well. Therefore, $x^\top w \geq \hat{x}^\top w$. We can conclude that:

$$0 < x^\top w + b \geq \hat{x}^\top w + b = BT(w) + b \tag{28}$$

which implies that $b > BT(w)$ as desired. □

From this point, we will assume $S_x = X$ without mentioning it explicitly.

3.3 AUXILIARY LEMMAS

We first define a special positive sample for every term. The special sample is a sample where all indices corresponding to the term will have positive values, and all other indexes will have negative values. The special samples will be used as the hardest positive samples to satisfy the MIN_+ property.

Definition 3.4. For a term $n \in [K]$, we define the special sample $x^{(n)} \in S_x$ of this term as follows:

$$\forall j \in A_n \; x_j^{(n)} = 1 \text{ and } \forall j \in [D] \setminus A_n \; x_j^{(n)} = -1 \tag{29}$$

We denote the set of all the special samples by $O = \{ x \in S_+ \mid \exists n \in [K] \; x = x^{(n)} \}$

Lemma 3.3. Given $\theta = (W, b)$, assume the following conditions are satisfied:

1. $\forall i \in [r], \forall j \in [D] \; w_{ij} \geq 0$.
2. For every $x \in O$ there exists $I \subseteq [r]$ such that $\sum_{i \in I} w_i \cdot x + b_i \geq 2$.

Then θ satisfies the MIN_+ property.

Proof. Let $x \in S_+$. Then $\exists n \in [K]$ such that $\forall j \in A_n \; x_j = 1$. By the second assumption $\exists I \subseteq [r]$ such that $\sum_{i \in I} w_i \cdot x^{(n)} + b_i \geq 2$.

For every $i \in [r]$ the following holds:

$$w_i \cdot x = \sum_{j \in [D]} w_{ij} x_j = \sum_{j \in A_n} w_{ij} + \sum_{j \in [D] \setminus A_n} x_j w_{ij} \tag{30}$$

From the first condition of the claim we can deduce that

$$\sum_{j \in A_n} w_{ij} + \sum_{j \in [D] \setminus A_n} x_j w_{ij} \geq \sum_{j \in A_n} w_{ij} \geq \sum_{j \in [D] \setminus A_n} w_{ij} = \sum_{j \in [D]} w_{ij} x_j^{(n)} = w_i \cdot x^{(n)} \tag{31}$$

Then:

$$\sum_{i \in I} (w_i \cdot x + b_i) \geq \sum_{i \in I} (w_i \cdot x^{(n)} + b_i) \geq 2 \tag{32}$$

and θ satisfies the MIN_+ property for x as required. □

The following definition will be very useful in our analysis.

Definition 3.5. Given a min-norm solution $\theta = (W, b)$, we say that the a solution $\hat{\theta} = (\hat{W}, \hat{b})$ is an i-modified solution if the following holds:

$$\forall i_0 \in [r] \setminus \{i\} \; \hat{w}_{i_0} = w_{i_0} \text{ and } \hat{b}_{i_0} = b_{i_0} \tag{33}$$

Thus, given a min-norm solution, to define an i-modified solution, we only need to define the neuron (w_i, b_i).

Lemma 3.4. Given a min-norm solution $\theta = (W, b)$, every $i \in [r]$ satisfies:

1. $b_i = BT(w_i)$.
2. $\forall j \in [D] \; w_{ij} \geq 0$.
3. $\exists n \in [K]$ such that $\forall j \in A_n \; w_{ij} \geq 0$ and $\forall j \in [D] \setminus A_n \; w_{ij} = 0$.

Proof. Property 1: Assume by contradiction that $b_i \neq BT(w_i)$. By Lemma 3.2, b_i has to be smaller than $BT(w_i)$, because otherwise θ is not a perfect solution. Now consider the i-modified solution, $\hat{\theta}$, which is defined by:

$$\hat{w}_i = w_i, \quad \hat{b}_i = BT(w_i) \tag{34}$$

By the assumption $\hat{b}_i > b_i$. Then, every $x \in S_x$ satisfies the following:

$$x \cdot w_i + b_i < x \cdot \hat{w}_i + \hat{b}_i \tag{35}$$

Since w satisfies the MIN$_+$ property, the above implies that $\hat{w}$ satisfies it as well.

From Definition 3.5 every $x \in S$ satisfies:

$$\forall i' \in [r] \setminus \{i\} \quad 0 > x \cdot w_{i'} + b_{i'} = x \cdot \hat{w}_{i'} + \hat{b}_{i'} \tag{36}$$

In addition, we saw in the proof of Lemma 3.2 that if $\hat{b}_i = BT(\hat{w})$ then $0 \geq x \cdot \hat{w}_i + \hat{b}_i$. Therefore, $\hat{w}$ satisfies the MIN property. By Lemma 3.1, $\hat{w}$ is a perfect solution.

From Definition 3.3 the bias threshold is nonpositive and therefore $b_i \leq \hat{b}_i \leq 0$ implies that $|\hat{b}_i| < |b_i|$. We know that $\hat{w}_i = w_i$ and therefore $\|(w_i, \hat{b}_i)\|_2^2 < \|(w_i, b_i)\|_2^2$ which contradicts the optimality of w.

Property 2: Assume by contradiction that $\exists j' \in [D]$ such that $w_{ij'} < 0$. Consider the following j-modified solution $\hat{w}$:

$$\forall j \in [D] \setminus \{j'\} \quad \hat{w}_{ij} = w_{ij} \quad \wedge \quad \hat{w}_{ij'} = 0 \quad \wedge \quad \hat{b}_i = BT(\hat{w}_i) \tag{37}$$

We want to show that:

$$\sum_{n \in [K]} V_n(w_i) = \sum_{n \in [K]} V_n(\hat{w}_i) \tag{38}$$

If $\exists n' \in [K]$ such that $j' \in A_{n'}$, then it follows that $V_{n'}(w_i) = V_{n'}(\hat{w}_i) = 0$ and Eq. (38) is satisfied. Otherwise, $j' \in A_{K+1}$, by Definition 3.1, the indices of A_{K+1} don't affect the value of the sums in Eq. (38). Therefore, this equation is satisfied in this case as well.

We know that $b_i = BT(w_i)$ according to Property 1 above, thus every $x \in S_x$ satisfies the following:

$$\begin{aligned} x \cdot w_i + b_i &= \sum_{j \in [D]} x_j w_{ij} + BT(w_i) \\ &= \sum_{j \in [D] \setminus \{j'\}} x_j w_{ij} + x_{j'} w_{ij'} - |w_{ij'}| - \sum_{j \in [D] \setminus \{j'\}} |w_{ij}| + \sum_{n \in [K]} 2V_n(w_i) \end{aligned} \tag{39}$$

We can see that $x_{j'} w_{ij'} - |w_{ij'}| \leq 0$. Then,

$$\begin{aligned} &\sum_{j \in [D] \setminus \{j'\}} x_j w_{ij} + x_{j'} w_{ij'} - |w_{ij'}| - \sum_{j \in [D] \setminus \{j'\}} |w_{ij}| + \sum_{n \in [K]} 2V_n(w_i) \\ &\leq \sum_{j \in [D] \setminus \{j'\}} x_j w_{ij} - \sum_{j \in [D] \setminus \{j'\}} |w_{ij}| + \sum_{n \in [K]} 2V_n(w_i) \\ &= \sum_{j \in [D]} x_j \hat{w}_{ij} - \|\hat{w}_i\|_1 + \sum_{n \in [K]} 2V_n(\hat{w}_i) = x \cdot \hat{w}_i + BT(\hat{w}_i) = x \cdot \hat{w}_i + \hat{b}_i \end{aligned} \tag{40}$$

Using the fact that $x \cdot w_i + b_i \leq x \cdot \hat{w}_i + \hat{b}_i$ with the fact that w satisfies the MIN$_+$ property, we can conclude that $\hat{w}$ satisfies this property too.

According to Property 1 above and Definition 3.5 we have:

$$\forall i' \in [r] \setminus \{i\} \quad BT(\hat{w}_{i'}) = BT(w_{i'}) = b_{i'} = \hat{b}_{i'} \tag{41}$$

In addition, we know that $\hat{b}_i = BT(\hat{w}_i)$ by Eq. (37). According to Lemma 3.2, we know that $\hat{w}$ is a perfect solution.

From Eq. (37), we know that $|w_{ij'}| > |\hat{w}_{ij'}| \Rightarrow \|w_i\|_1 > \|\hat{w}_i\|_1$. Combining this with Eq. (38) and the fact that the bias threshold is nonpositive we can conclude that

$$\|w_i\|_1 + 2\sum_{n \in [K]} V_n(w_i) < \|\hat{w}_i\|_1 + 2\sum_{n \in [K]} V_n(\hat{w}_i) \Rightarrow |BT(w_i)| > |BT(\hat{w}_i)| \Rightarrow |b_i| > |\hat{b}_i| \tag{42}$$

Therefore $\|(\hat{w}_i;\hat{b}_i)\|_2^2 < \|(w_i;b_i)\|_2^2$ in contradiction to the optimality of $(w_i;b_i)$.

Property 3: Assume by contradiction that there exist $i \in [r]$ such that:

$$\exists n_1 \neq n_2 \in [K+1] \text{ such that } \exists j \in A_{n_1}\ w_{ij} > 0 \text{ and } \exists j \in A_{n_2}\ w_{ij} > 0 \tag{43}$$

Without loss of generality we assume:

$$\sum_{j \in A_{n_1}} w_{ij} \geq \sum_{j \in A_{n_2}} w_{ij} \tag{44}$$

Let's look on the following $\hat{}$-modified $\hat{}$ which is defined by:

$$\forall j \in A_{n_2}\ \hat{w}_{ij} = 0 \text{ and } \forall j \in [D] \setminus A_{n_2}\ \hat{w}_{ij} = w_{ij} \text{ and } \hat{b}_i = BT(\hat{w}_i) \tag{45}$$

First, we will show that $\hat{b}_i \leq b_i$. If $n_2 \neq K+1$, from Definition 3.3 and the assumption that $|A_{n_2}| > 1$, the following holds:

$$\hat{b}_i = BT(\hat{w}_i) = BT(w_i) + \sum_{j \in A_{n_2}} |w_{ij}| \leq 2V_{n_2}(w_i) = b_i + \sum_{j \in A_{n_2}} w_{ij} \leq 2V_{n_2}(w_i) \leq b_i \tag{46}$$

Otherwise, $n_2 = K+1$ and from Definition 3.3 the following holds:

$$\hat{b}_i = BT(\hat{w}_i) = BT(w_i) + \sum_{j \in A_{n_2}} |w_{ij}| = b_i + \sum_{j \in A_{n_2}} w_{ij} \leq b_i \tag{47}$$

In both cases $\hat{b}_i \leq b_i$ as required.

Given $n \in [K]$, we know that $(w;b)$ is a perfect solution and thus it satisfies the MIN_+ property. Then, for $x^{(n)}$ there exists $I \subseteq [r]$ such that:

$$\sum_{i' \in I} w_{i'} \cdot x^{(n)} + b_{i'} \leq 2 \tag{48}$$

We will show that there exists $I' \subseteq [r]$ such that:

$$\sum_{i' \in I'} \hat{w}_{i'} \cdot x^{(n)} + \hat{b}_{i'} \leq 2 \tag{49}$$

Recall, by Definition 3.5, for any $i' \in [r] \setminus \{i\}$ we know that $\hat{w}_{i'} \cdot x^{(n)} + \hat{b}_{i'} = w_{i'} \cdot x^{(n)} + b_{i'}$.

If $n \neq n_2$, due to Property 1 and Property 2 above and the fact that $\hat{b}_i \leq b_i$ the following holds:

$$\begin{aligned} w_i \cdot x^{(n)} + b_i &= \sum_{j \in [D] \setminus A_{n_2}} w_{ij} x_j^{(n)} + \sum_{j \in A_{n_2}} w_{ij} + b_i < \sum_{j \in [D] \setminus A_{n_2}} w_{ij} x_j^{(n)} + b_i \leq \sum_{j \in [D]} \hat{w}_{ij} x_j^{(n)} + \hat{b}_i \\ &= \hat{w}_i \cdot x^{(n)} + \hat{b}_i \end{aligned} \tag{50}$$

Therefore,

$$\sum_{i' \in I} \hat{w}_{i'} \cdot x^{(n)} + \hat{b}_{i'} \leq \sum_{i' \in I} w_{i'} \cdot x^{(n)} + b_{i'} \leq 2 \tag{51}$$

Otherwise, $n = n_2$. By the fact that $b_i = BT(w_i) \geq 0$, Property 2 above and Eq. (44) the following holds:

$$w_i \cdot x^{(n)} + b_i = \sum_{j \in A_{n_2}} w_{ij} \leq \sum_{j \in [D] \setminus A_{n_2}} w_{ij} + b_i \leq \sum_{j \in A_{n_2}} w_{ij} \leq \sum_{j \in A_{n_1}} w_{ij} \leq \sum_{j \in A_{n_2}} w_{ij} + b_i \leq \sum_{j \in A_{n_1}} w_{ij} \leq \sum_{j \in A_{n_2}} w_{ij} \leq 0 \tag{52}$$

Therefore, using Definition 3.5, $\hat{}$ satisfies the following:

$$\sum_{i' \in I \setminus \{i\}} \hat{w}_{i'} \cdot x^{(n)} + \hat{b}_{i'} = \sum_{i' \in I \setminus \{i\}} w_{i'} \cdot x^{(n)} + b_{i'} \leq \sum_{i' \in I} w_{i'} \cdot x^{(n)} + b_{i'} \leq 2 \tag{53}$$

We can conclude that $\hat{}$ satisfies Eq. (49). Combining this with Property 2 above, we can see that $\hat{}$ meets the condition of Lemma 3.3 and then it satisfies the MIN_+ property.

According to Property 1 above and De nition 3.5:

$$\forall i' \in [r] \setminus \{i\}\ BT(\hat{w}_{i'}) = BT(w_{i'}) = b_{i'} = \hat{b}_{i'} \tag{54}$$

In addition, we know that $\hat{b}_i = BT(\hat{w}_i)$ by Eq. (45). According to Lemma 3.2, the solution $\hat{}$ is a perfect solution.

As we saw $0 \quad \hat{b}_i \quad b_i < \hat{b}_i \quad |b_i|, \forall j \in A_{n_2}\ w_{ij} \quad 0 = \hat{w}_{ij}$ and $\exists j \in A_{n_2}\ w_{ij} > 0 = \hat{w}_{ij}$ and therefore $\|(w_i; b_i)\|_2^2 > \|(\hat{w}_i; \hat{b}_i)\|_2^2$ which contradicts the optimality of . □

3.4 ALIGNMENT LEMMAS

The following three lemmas show the alignment properties of the min-norm solution.

Lemma 3.5. Given a min-norm solution $= (W; b)$, every $i \in [r]$ either aligns with some term $m \in [K]$ or it holds that $w_i = 0; b_i = 0$

Proof. Given $i \in [r]$, if $w_i = 0$ the claim is true by Property 1 of Lemma 3.4. Otherwise, by Property 3 of Lemma 3.4, $\exists n \in [K]$ such that:

$$\forall j \in A_n\ w_{ij} \quad 0 \text{ and } \exists j \in A_n\ w_{ij} > 0 \text{ and } \forall j \in [D] \setminus A_n\ w_{ij} = 0 \tag{55}$$

Assume by contradiction that:

$$\exists j_1, j_2 \in A_n \text{ such that } w_{ij_1} \neq w_{ij_2} \tag{56}$$

Without loss of generality, we assume $w_{ij_1} > w_{ij_2}$ and $w_{ij_2} = \min_{j \in A_n} \{w_{ij}\}$.

De ne the following i-modi ed solution $\hat{}$:

$$\forall j \in A_n\ \hat{w}_{ij} = w_{ij_2} \text{ and } \forall j \in [D] \setminus A_n\ \hat{w}_{ij} = w_{ij} \text{ and } \hat{b}_i = BT(\hat{w}_i) \tag{57}$$

Note that $V_n(w_i) = V_n(\hat{w}_i) = w_{ij_2}$.

Given $x \in S_+$, we know that satis es the MIN property. Then $\exists l$ [r] such that:

$$\sum_{i' \in I} w_{i'} \quad x + b_{i'} \quad 2 \tag{58}$$

Recall, by De nition 3.5, for any $i' \in [r] \setminus \{i\}$ we know that $\hat{w}_{i'} \quad x + \hat{b}_{i'} = w_{i'} \quad x + b_{i'}$

If $x \quad t_n = \|t_n\|_1$, due to Property 3 of Lemma 3.4 the following holds:

$$\begin{aligned} w_i \quad x + b_i &= \sum_{j \in A_n} w_{ij} + BT(w_i) = \sum_{j \in A_n} w_{ij} \sum_{j \in A_n} w_{ij} + 2V_n(w_i) = 2w_{ij_2} \\ &= \sum_{j \in A_n} \hat{w}_{ij} \sum_{j \in A_n} \hat{w}_{ij} + 2w_{ij_2} = \sum_{j \in A_n} \hat{w}_{ij} \|\hat{w}\|_1 + 2V_n(\hat{w}_i) = \hat{w}_i \quad x + \hat{b}_i \end{aligned} \tag{59}$$

Then we can conclude:

$$\sum_{i' \in I} \hat{w}_{i'} \quad x + \hat{b}_{i'} = \sum_{i' \in I} w_{i'} \quad x + b_{i'} \quad 2 \tag{60}$$

Otherwise, $x \quad t_n < \|t_n\|_1$ and by De nition 3.3:

$$w_i \quad x + b_i = \sum_{j \in A_n} w_{ij} x_j + BT(w_i) \sum_{j \in A_n \setminus \{j_2\}} w_{ij} \quad w_{ij_2} \sum_{j \in A_n \setminus \{j_2\}} w_{ij} \quad w_{ij_2} + 2w_{ij_2} = 0 \tag{61}$$

Therefore,

$$\sum_{i' \in I \setminus \{i\}} \hat{w}_{i'} \quad x + \hat{b}_{i'} = \sum_{i' \in I \setminus \{i\}} w_{i'} \quad x + b_{i'} \sum_{i' \in I} w_{i'} \quad x + b_{i'} \quad 2 \tag{62}$$

We can conclude that $\hat{}$ satisfies the MIN$_+$ property.

According to Property 1 of Lemma 3.4 and Definition 3.5:

$$\forall i' \in [r] \setminus \{i\}\ BT(\hat{w}_{i'}) = BT(w_{i'}) = b_{i'} = \hat{b}_{i'} \tag{63}$$

In addition, we know that $\hat{b}_i = BT(\hat{w}_i)$ by Eq. (57). According to Lemma 3.2, the solution is a perfect solution.

Finally, we can see that $\|w_i\|_1 > \|\hat{w}_i\|_1$ implies that $|BT(w_i)| > |BT(\hat{w}_i)|$ and $\forall j \in [D] : w_{ij} \neq \hat{w}_{ij}$. Thus, we have $\|(w_i; b_i)\|_2^2 > \|(\hat{w}_i; \hat{b}_i)\|_2^2$. This is in contradiction to the optimally of , as desired.

We can now define $_i = V_n(w_i)$ and we know that the neuron satisfies:

$$\forall j \in A_n\ w_{ij} = {}_i \text{ and } \forall j \in [D] \setminus A_n\ w_{ij} = 0 \tag{64}$$

Therefore, $w_i = {}_i t_n$; $b_i = {}_i(2 - \|t_n\|_1)$ and we can say that neuron aligns the term . □

Lemma 3.6. Given a min-norm solution $= (W ; b)$, every 2 neurons $i_1; i_2 \in [r]$ that align with term $n \in [K]$ satisfy $_{i_1} = {}_{i_2}$.[1]

Proof. Given $i_1; i_2 \in [r]$ that align with term $n \in [K]$ we have $w_{i_1} = {}_{i_1} t_n$ and $w_{i_2} = {}_{i_2} t_n$. Assume by contradiction that $_{i_2} \neq {}_{i_1}$. Define the following solution $\hat{}$:

$$\begin{aligned} &\forall i \in [r] \setminus \{i_1; i_2\}\ \hat{w}_i = w\ ;\ \hat{b}_i = b_i \\ &\hat{w}_{i_1} = \frac{{}_{i_1} + {}_{i_2}}{2} t_n;\ \hat{b}_{i_1} = BT(\hat{w}_{i_1}) \\ &\hat{w}_{i_2} = \frac{{}_{i_1} + {}_{i_2}}{2} t_n;\ \hat{b}_{i_2} = BT(\hat{w}_{i_2}) \end{aligned} \tag{65}$$

Note, $\forall i' \in [r] \setminus \{i_1; i_2\}$ we know that $\hat{w}_{i'}\ x + \hat{b}_{i'} = w_{i'}\ x + b_{i'}$

Given $x \in S_+$, we know that satisfies the MIN$_+$ property. Then, $\exists I \subseteq [r]$ such that:

$$\sum_{i' \in I} w_{i'}\ x + b_{i'} \quad 2 \tag{66}$$

If $x\ t_n = \|t_n\|_1$ we can calculate the following:

$$w_{i_1}\ x + b_{i_1} + w_{i_2}\ x + b_{i_2} = 2\ {}_{i_1} + 2\ {}_{i_2} = \hat{w}_{i_1}\ x + \hat{b}_{i_1} + \hat{w}_{i_2}\ x + \hat{b}_{i_2} \tag{67}$$

Therefore,

$$\sum_{i' \in I} \hat{w}_{i'}\ x + \hat{b}_{i'} = \sum_{i' \in I} w_{i'}\ x + b_{i'} \quad 2 \tag{68}$$

Otherwise, $x\ t_n < \|t_n\|_1$ thus $w_{i_1}\ x + b_{i_1} \quad 0, w_{i_2}\ x + b_{i_2} \quad 0$. Then,

$$\sum_{i' \in I \setminus \{i_1; i_2\}} \hat{w}_{i'}\ x + \hat{b}_{i'} = \sum_{i' \in I \setminus \{i_1; i_2\}} w_{i'}\ x + b_{i'} \quad \sum_{i' \in I} w_{i'}\ x + b_{i'} \quad 2 \tag{69}$$

Combining this with Property 2 of Lemma 3.4 we can see that $\hat{}$ meets the condition of Lemma 3.3 and then it satisfies the MIN$_+$ property.

According to Property 1 of Lemma 3.4:

$$\forall i' \in [r] \setminus \{i_1; i_2\}\ BT(\hat{w}_{i'}) = BT(w_{i'}) = b_{i'} = \hat{b}_{i'} \tag{70}$$

In addition, we know that $\hat{b}_{i_1} = BT(\hat{w}_{i_1})$ and $\hat{b}_{i_2} = BT(\hat{w}_{i_2})$ by Eq. (65). According to Lemma 3.2, the solution is a perfect solution.

[1] $_i$ were defined in the proof of the previous lemma.

We will prove that $\sum_{i\in[r]} \|(w_i; b_i)\|_2^2 > \sum_{i\in[r]} \|(\hat{w}_i; \hat{b}_i)\|_2^2$. This will contradict the optimality of $(w_i; b_i)$. Note that $\|(w_{i_1}; b_{i_1})\|_2^2 = \gamma_{i_1}^2 \|A_n\| + \gamma_{i_1}^2 (\|A_n\| - 2)^2$. Then:

$$\begin{aligned}
\sum_{i\in[r]} \|(w_i; b_i)\|_2^2 - \sum_{i\in[r]} \|(\hat{w}_i; \hat{b}_i)\|_2^2 &= \|(w_{i_1}; b_{i_1})\|_2^2 + \|(w_{i_2}; b_{i_2})\|_2^2 - \|(\hat{w}_{i_1}; \hat{b}_{i_1})\|_2^2 - \|(\hat{w}_{i_2}; \hat{b}_{i_2})\|_2^2 \\
&= \gamma_{i_1}^2 + \gamma_{i_2}^2 - 2\left(\frac{\gamma_{i_1} + \gamma_{i_2}}{2}\right)^2 \|A_n\| + (\|A_n\| - 2)^2 \\
&= \frac{1}{2}\gamma_{i_1}^2 - \gamma_{i_1}\gamma_{i_2} + \frac{1}{2}\gamma_{i_2}^2 \|A_n\| + (\|A_n\| - 2)^2 \\
&= \left(\frac{1}{\sqrt{2}}\gamma_{i_1} - \frac{1}{\sqrt{2}}\gamma_{i_2}\right)^2 \|A_n\| + (\|A_n\| - 2)^2
\end{aligned}$$

For $\gamma_{i_1} \neq \gamma_{i_2}$, we get $\sum_{i\in[r]} \|(w_i; b_i)\|_2^2 - \sum_{i\in[r]} \|(\hat{w}_i; \hat{b}_i)\|_2^2 > 0$, as needed. □

Lemma 3.7. Given a min-norm solution $\theta = (W, b)$, if $I \subseteq [r]$ is the set of all neurons that align with term $n \in [K]$, then $\sum_{i\in S} \gamma_i = 1$.

Proof. Assume by contradiction that $\sum_{i\in I} \gamma_i \neq 1$. If $\sum_{i\in I} \gamma_i < 1$, then $\forall i^0 \in [r]\setminus I$, by Lemma 3.5 we know that the neuron i^0 aligns with another term or is equal to 0, then $x^{(n)} \cdot w_i \leq 0$. By Property 1 of Lemma 3.4, $b_{i^0} = -BT(w_{i^0}) \leq 0$. Therefore, $x^{(n)} \cdot w_{i^0} + b_{i^0} \leq 0$. Then, for every $i^0 \in [r]$ the following holds:

$$\sum_{i\in I^0} w_i \cdot x^{(n)} + b_i \leq \sum_{i\in I^0 \setminus I} w_i \cdot x^{(n)} + b_i = \sum_{i\in I^0 \setminus I} \|A_n\| \gamma_i + (2 - \|A_n\|)\gamma_i = 2\sum_{i\in I^0 \setminus I} \gamma_i < 2 \tag{71}$$

and thus θ doesn't satisfy the MIN-1 property. By Lemma 3.1, this contradicts the fact that θ is a perfect solution.

If $\sum_{i\in I} \gamma_i > 1$, we choose an arbitrary $\hat{i} \in I$. Define $\hat{\theta}$ as follows:

$$\begin{aligned}
&\forall i \in [r]\setminus I \quad \hat{w}_i = w_i, \ \hat{b}_i = b_i \\
&\forall i \in I\setminus\{\hat{i}\} \quad \hat{w}_i = 0, \ \hat{b}_i = 0 \\
&\hat{w}_{\hat{i}} = t_n, \ \hat{b}_{\hat{i}} = -BT(\hat{w}_{\hat{i}})
\end{aligned} \tag{72}$$

Given $x^{(\theta)} \in O$, we know that θ satisfies the MIN-1. Thus, $\exists i^0 \in [r]$ such that:

$$\sum_{i\in I^0} w_{i^0} \cdot x^{(\theta)} + b_i \geq 2 \tag{73}$$

We will show that there exists $\hat{i} \in [r]$ such that:

$$\sum_{i^0 \in \hat{I}} \hat{w}_{i^0} \cdot x^{(\theta)} + \hat{b}_{i^0} \geq 2 \tag{74}$$

If $\theta = n$, then:

$$\hat{w}_{\hat{i}} \cdot x^{(\theta)} + \hat{b}_{\hat{i}} = t_n \cdot x^{(\theta)} - BT(t_n) = \|t_n\|_1 - \|t_n\|_1 + 2 = 2 \tag{75}$$

Therefore, by choosing $\hat{I} = \{\hat{i}\}$, we can show that $\hat{\theta}$ satisfies Eq. (74).

Otherwise, $\theta \neq n$ and then every $i \in I$ satisfies:

$$w_i \cdot x^{(\theta)} + b_i = \gamma_i \|A_n\| - \gamma_i(\|A_n\| - 2) < 0 \tag{76}$$

From Eq. (72) we can conclude:

$$\sum_{i\in I^0\setminus I} \hat{w}_i \cdot x^{(\theta)} + \hat{b}_i = \sum_{i\in I^0\setminus I} w_i \cdot x^{(\theta)} + b_i \geq \sum_{i\in I^0} w_i \cdot x^{(\theta)} + b_i \geq 2 \tag{77}$$

Therefore, $\hat{\theta}$ satisfies Eq. (74) for $x^{(n)}$. In addition, by Property 2 of Lemma 3.4 and Eq. (72), we know that all the weights of the neurons in $\hat{\theta}$ are nonnegative. Then, $\hat{\theta}$ meets the condition of Lemma 3.3 and it satisfies the XOR property.

According to Property 1 of Lemma 3.4 and Eq. (72):

$$\forall i \in [r] \cup I \quad BT(\hat{w}_i) = BT(w_i) = b_i = \hat{b}_i \tag{78}$$

In addition, we know that $\forall i \in Inf\hat{\theta} \quad \hat{b}_i = 0 = BT(\hat{w}_i)$ and $\hat{b}_{\hat{i}} = BT(\hat{w}_{\hat{i}})$. Therefore, according to Lemma 3.2, the solution $\hat{\theta}$ is a perfect solution.

We will prove that $\sum_{i\in[r]} \|(w_i; b_i)\|_2^2 > \sum_{i\in[r]} \|(\hat{w}_i; \hat{b}_i)\|_2^2$. This will contradict the optimality of θ. Indeed:

$$\sum_{i\in[r]} \|(w_i; b_i)\|_2^2 - \sum_{i\in[r]} \|(\hat{w}_i; \hat{b}_i)\|_2^2 = \sum_{i\in I} \|(w_i; b_i)\|_2^2 - \|(\hat{w}_{\hat{i}}; \hat{b}_{\hat{i}})\|_2^2 = \tag{79}$$

$$\sum_{i\in I} \alpha_i^2 |A_n| + \alpha_i^2 (|A_n| - 2)^2 - |A_n| - (|A_n| - 2)^2 =$$

$$\sum_{i\in I} \left(\alpha_i^2 - 1\right) |A_n| + (|A_n| - 2)^2$$

Since $\sum_{i\in I} \alpha_i > 1$, we have $\sum_{i\in[r]} \|(w_i; b_i)\|_2^2 - \sum_{i\in[r]} \|(\hat{w}_i; \hat{b}_i)\|_2^2 > 0$, which completes the proof.

□

3.5 FINISHING THE PROOF OF THEOREM 6.1

Proof. Given a min-norm solution $\theta = (W; b)$, by Lemma 3.5, each neuron $i \in [r]$ aligns with some term $n_i \in [K]$ or it is equal to 0. Assume by contradiction that there exists a term $n \in [K]$ that is not aligned, namely $\forall i \in [r] \; n_i \neq n$. Consider the special positive sample $x^{(n)} \in S_+$. From the definition of $x^{(n)}$, every $j \in [D] \setminus A_n$ satisfies $x_j^{(n)} = -1$. Then,

$$\forall i \in [r] \quad x^{(n)} w_i = \alpha_n x^{(n)} t_{n_i} = \alpha_n \|t_{n_i}\|_1 < 0 \tag{80}$$

By the first property of Lemma 3.5, $\forall i \in [r]; \; b_i = BT(w_i) \geq 0$. Therefore, $\forall i \in [r] : x^{(n)} w_i + b_i < 0$, in contradiction to the fact that θ is a perfect solution.

In conclusion, every term $n \in [K]$ aligns with a set of neurons $I \subseteq [r]$. By Lemma 3.7, we know that $\sum_{i\in I} \alpha_i = 1$. By Lemma 3.6, we know that every two neurons i_1, i_2 that align with the same term satisfy $v_{i_1} = v_{i_2}$. Therefore, θ is a DNF recovery solution. □

4 EXPERIMENT DETAILS AND ADDITIONAL RESULTS

Selected read-once DNFs: In this section we provide additional experiments for the following types of read-once DNFs.

- f_1 - 3-term read-once DNF where the length of every term is 3 for $D = 9$
- f_2 - 4-term read-once DNF where the length of the terms are 4,5,5,6 for $D = 25$
- f_3 - 8-term read-once DNF where the length of the terms are 3,3,4,4,4,4,5,5 for $D = 40$
- f_4 - 10-term read-once DNF where the length of the terms are 3,3,4,4,4,4,6,6,6,6 for $D = 50$
- f_5 - 15-term read-once DNF where of every term is 5 for $D = 100$

General details: In all the experiments, "small initialization" refers to initializing weights from $W^{(0)} \sim N(0; 10^{-6})$ and $b^{(0)} = [0]^D$. The learning rate for SGD is $= 10^{-3}$, the number of hidden units is $= 2000$ and the batch size is 32. We create the train set by sampling uniformly from $[-1, 1]^D$. All the experiments can run on any single GPU. Training a single network can take up to two hours.

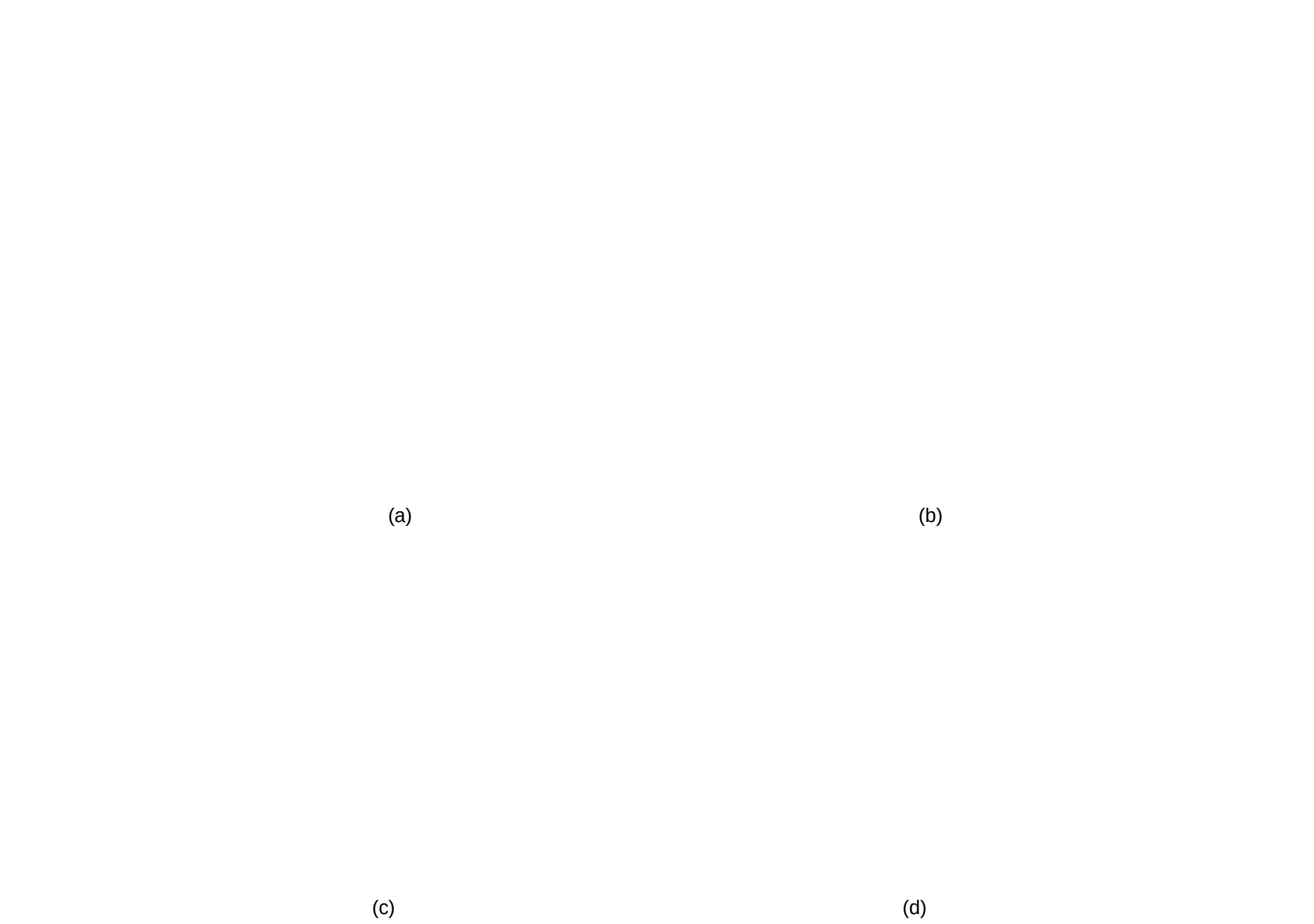

Figure 5: Test accuracy for the convex network with small initialization, convex network with large initialization and standard networks. Figure (a) shows the performance when learning f_1, Figure (b) for f_3, Figure (c) for f_4 and Figure (d) for f_3 (Results for f_2 were presented in the main paper).

Weight Matrix Visualization: When presenting weight matrices, we first cluster the neurons using the Hierarchical clustering algorithm[2]. We then plot the weight values in an image, where neurons clustered together appear in consecutive rows. Note this of course does not change the model itself, but makes it easy to see if there are well clustered neurons (as in the DNF recovery case.

Sample Complexity Experiments: We evaluate test accuracy as a function of the training sample size for different models. Results are shown in Figure 5. Specifically, we compare the convex network with small initialization (see details above), the convex network with large initialization (we take $W^{(0)} \sim N(0, 1)$), and a "standard" network with one hidden layer, same width as the convex network and Xavier initialization (we checked different initialization schemes, including small Gaussian initialization, and verified that this not affect the results). We run every experiment 10 times and present the mean performance and the std of this mean (the std is small and smaller than the line width).

Implementation of the Statistical Queries (SQ) Methods: In Figure 2a of the main paper, we present results of the statistical query algorithm. We implemented the algorithm described in Mansour & Schain (2001). We view the tolerance as a hyperparameter. Therefore, we use 10% of the train set as validation for tuning. We present the performance of the SQ algorithm only for $D = 9$, because for larger dimension the algorithm failed in creating a DNF for the range of train set sizes tested.

[2] We used scipy.cluster.hierarchy.linkage. with centroid as a method.

Figure 6: (a-c) Effect of training size on the learned model (see learned models for different training sizesin the text) for the ground-truth model f_1 . Panels a-c correspond to training sizes 800,1500 and 7500. (d) Result for training on 7500 with large initialization.

Figure 7: Same as Figure 6 but with f_2 as ground truth.

Figure 8: Same as Figure 6 but with f_3 as ground truth.

Figure 9: Same as Figure 6 but with f_4 as ground truth.

Algorithm 1 Reconstruction Procedure

Input: Network $\boldsymbol{\theta} = (\boldsymbol{W}, \boldsymbol{b}, c)$, DNF f^*, fixed sets $A, B \subseteq [0,1]^L$.
Output: True if the network with parameter $\boldsymbol{\theta}$ reconstructs DNF f^*, False otherwise.

```
for (a, b) ∈ A × B do
    W' ← [0]^{rD}                          ▷ W' will be a {0,1} matrix where each row corresponds to a term in a DNF.
    for 1 ≤ i ≤ r do
        if ℓ∞(w_i) ≥ a * ℓ∞(W) then        ▷ Taking into account only meaningful neurons
            for 1 ≤ j ≤ r do
                if w_ij ≥ ℓ∞(w_i) * b then   ▷ Taking into account only meaningful values
                    w'_ij ← 1
                end if
            end for
        end if
    end for
    if The set of terms represented by W' is exactly the set of terms of the DNF f* then
        return True
    end if
end for
return False
```

Learned models for different training sizes: In the main paper, we show empirically that learning convex networks with small init and GD leads to a DNF recovery solution, and we also show formally that in the population risk DNF recovery is norm minimizing. Here we show explicit model weights for different training sizes, demonstrating that approximate DNF recovery solutions are obtained for fairly small sample sizes. Figures 6,7,8,9,10 (panels a-c) show these results. In panel d of these figures we show the learned model for when learning with large Gaussian initialization and with the same train set size. It can be seen that larger initialization does not results in the recovery-DNF solution (note we are also visualizing these solutions using clustering as explained above, and there is clearly no cluster structure in the solution).

DNF reconstruction: In Figure 2b of the main paper we present accuracy results for DNF reconstruction. To obtain these, we take the learned model and apply a simple rounding procedure to check if this model reconstructs the ground-truth DNF. The procedure is outlined in Algorithm 1. In the procedure, we create a $\{0,1\}$ matrix $\boldsymbol{W}'$ where the column indices of 1s in each row correspond to a term of a DNF. Thus, $\boldsymbol{W}'$ represents a set of terms. If the set of terms of $\boldsymbol{W}'$ is exactly the set of terms of the input DNF, the procedure returns True. In our experiments we ran the procedure with inputs $A = [0, 0.1, 0.2, \ldots, 0.9]$ and $B = [0, 0.2, 0.4, \ldots, 0.8]$.

The effect of learning c: Figure 11 shows that fixing the learnable parameter c to -1 does not effect the structure of the solution. In this experiment, we took 2 networks with the same width: One with learnable c initialized to 0, and the second with fixed $c = -1$. We initialize the other weights with the same values, and train them with the same train set for the same number of steps. Finally we plot the solution that the network learns. We can say they both recover the underlying DNF.

Tabular datasets: We consider the three UCI datasets: kr-vs-kp, Splice, and diabetes. For these, we convert the input into binary by changing categorical variables to one-hot. We also consider binary classification such that in kr-vs-kp the class 'won' is positive considered and 'notwon' is negative, in Splice the classes 'EI' and 'IE' are considered positive and 'N' negative, and diabetes is binary by design. We train on 90% of the data and test on 10%. The reconstruction process is identical to algorithm 1 when instead of validate if $\boldsymbol{W}'$ is identical to f^*, we return the $\boldsymbol{W}'$ with the best accuracy on the train set.

The relevant code can be found in our repository: `https://github.com/idobronstein/Exploring-the-Inductive-Bias-of-Neural-Networks-for-Learning-Read-once-DNFs.git`.