Why do Larger Models Generalize Better? A Theoretical Perspective via the XOR Problem
Supplementary Material

Contents

A Experiment in Figure 1

We tested the generalization performance in the setup of Section 4. We considered networks with number of channels 4,6,8,20,50,100 and 200. The distribution in this setting has $p_+ = 0.5$ and $p_- = 0.9$ and the training sets are of size 12 (6 positive, 6 negative). Note that in this case the training set contains non-diverse points with high probability. The ground truth network can be realized by a network with 4 channels. For each number of channels we trained a convolutional network 100 times and averaged the results. In each run we sampled a new training set and new initialization of the weights according to a gaussian distribution with mean 0 and standard deviation

0.00001. For each number of channels c, we ran gradient descent with learning rate $\frac{0.04}{c}$ and stopped it if it did not improve the cost for 20 consecutive iterations or if it reached 30000 iterations. The last iteration was taken for the calculations. We plot both average test error over all 100 runs and average test error only over the runs that ended at 0% train error. In this case, for each number of channels 4,6,8,20,50,100,200 the number of runs in which gradient descent converged to a 0% train error solution is 62, 79, 94, 100, 100, 100, 100, respectively.

B Proofs for Section 3

In the XOR problem, we are given a training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^4 \subseteq \{\pm 1\}^2 \times \{\pm 1\}^2$ consisting of points $\boldsymbol{x}_1 = (1,1)$, $\boldsymbol{x}_2 = (-1,1)$, $\boldsymbol{x}_3 = (-1,-1)$, $\boldsymbol{x}_4 = (1,-1)$ with labels $y_1 = 1$, $y_2 = -1$, $y_3 = 1$ and $y_4 = -1$, respectively. Our goal is to learn the XOR function $f^* : \{\pm 1\}^2 \to \{\pm 1\}$, such that $f^*(\boldsymbol{x}_i) = y_i$ for $1 \le i \le 4$, with a neural network and gradient descent.

Neural Architecture: For this task we consider the following two-layer fully connected network.

$$N_W(\boldsymbol{x}) = \sum_{i=1}^{k} \left[\sigma\left(\boldsymbol{w}^{(i)} \cdot \boldsymbol{x}\right) - \sigma\left(\boldsymbol{u}^{(i)} \cdot \boldsymbol{x}\right)\right] \tag{1}$$

where $W \in \mathbb{R}^{2k \times 2}$ is the weight matrix whose rows are the $\boldsymbol{w}^{(i)}$ vectors followed by the $\boldsymbol{u}^{(i)}$ vectors, and $\sigma(x) = \max\{0, x\}$ is the ReLU activation applied element-wise. We note that f^* can be implemented with this network for $k = 2$ and this is the minimal k for which this is possible. Thus we refer to $k > 2$ as the overparameterized case.

Training Algorithm: The parameters of the network $N_W(\boldsymbol{x})$ are learned using gradient descent on the hinge loss objective. We use a constant learning rate $\eta = \frac{c_\eta}{k}$, where $c_\eta < \frac{1}{2}$. The parameters N_W are initialized as IID Gaussians with zero mean and standard deviation $\sigma_g \le \frac{c_\eta}{16k^{3/2}}$. We consider the hinge-loss objective:

$$\ell(W) = \sum_{(\boldsymbol{x},y)\in S} \max\{1 - yN_W(\boldsymbol{x}), 0\}$$

where optimization is only over the first layer of the network. We note that for $k \ge 2$ any global minimum W of ℓ satisfies $\ell(W) = 0$ and $\text{sign}(N_W(\boldsymbol{x}_i)) = f^*(\boldsymbol{x}_i)$ for $1 \le i \le 4$.

Notations: We will need the following notations. Let W_t be the weight matrix at iteration t of gradient descent. For $1 \le i \le k$, denote by $\boldsymbol{w}_t^{(i)} \in \mathbb{R}^2$ the i^{th} weight vector at iteration t. Similarly we define $\boldsymbol{u}_t^{(i)} \in \mathbb{R}^2$ to be the $k+i$ weight vector at iteration t. For each point $\boldsymbol{x}_i \in S$ define the following sets of neurons:

$$W_t^+(i) = \left\{j \mid \boldsymbol{w}_t^{(j)} \cdot \boldsymbol{x}_i > 0\right\}$$

$$U_t^+(i) = \left\{j \mid \boldsymbol{u}_t^{(j)} \cdot \boldsymbol{x}_i > 0\right\}$$

and for each iteration t, let $a_i(t)$ be the number of iterations $0 \le t' \le t$ such that $y_i N_{W_{t'}}(\boldsymbol{x}_i) < 1$.

B.1 Overparameterized Network

Lemma B.1. ***Exploration at initialization.*** *With probability at least* $1 - 8e^{-8}$*, for all* $1 \le j \le 4$

$$\frac{k}{2} - 2\sqrt{k} \le \left|W_0^+(j)\right|, \left|U_0^+(j)\right| \le \frac{k}{2} + 2\sqrt{k}$$

Proof. Without loss of generality consider $\left|W_0^+(1)\right|$. Since the sign of a one dimensional Gaussian random variable is a Bernoulli random variable, we get by Hoeffding's inequality

$$\mathbb{P}\left(\left|\left|W_0^+(1)\right| - \frac{k}{2}\right| < 2\sqrt{k}\right) \leq 2e^{-\frac{2(2^2k)}{k}} = 2e^{-8}$$

Since $\left|W_0^+(1)\right| + \left|W_0^+(3)\right| = k$ with probability 1, we get that if $\left|\left|W_0^+(1)\right| - \frac{k}{2}\right| < 2\sqrt{k}$ then $\left|\left|W_0^+(3)\right| - \frac{k}{2}\right| < 2\sqrt{k}$. The result now follows by symmetry and the union bound. □

Lemma B.2. *With probability* $\geq 1 - \frac{\sqrt{8k}}{\sqrt{\pi}e^{8k}}$*, for all* $1 \leq j \leq k$ *and* $1 \leq i \leq 4$ *it holds that* $\left|\boldsymbol{w}_0^{(j)} \cdot \boldsymbol{x}_i\right| \leq \frac{\sqrt{2}\eta}{4}$ *and* $\left|\boldsymbol{u}_0^{(j)} \cdot \boldsymbol{x}_i\right| \leq \frac{\sqrt{2}\eta}{4}$.

Proof. Let Z be a random variable distributed as $\mathcal{N}(0, \sigma^2)$. Then by Proposition 2.1.2 in Vershynin (2017), we have

$$\mathbb{P}\left[|Z| \geq t\right] \leq \frac{2\sigma}{\sqrt{2\pi}t} e^{-\frac{t^2}{2\sigma^2}}$$

Therefore, for all $1 \leq j \leq k$ and $1 \leq i \leq 4$,

$$\mathbb{P}\left[\left|\boldsymbol{w}_0^{(j)} \cdot \boldsymbol{x}_i\right| \geq \frac{\sqrt{2}\eta}{4}\right] \leq \frac{1}{\sqrt{8\pi k}} e^{-8k}$$

and

$$\mathbb{P}\left[\left|\boldsymbol{u}_0^{(j)} \cdot \boldsymbol{x}_i\right| \geq \frac{\sqrt{2}\eta}{4}\right] \leq \frac{1}{\sqrt{8\pi k}} e^{-8k}$$

The result follows by applying a union bound over all $2k$ weight vectors and the four points $\boldsymbol{x}_i$, $1 \leq i \leq 4$. □

Lemma B.3. ***Clustering Dynamics. Lemma 3.2 restated and extended.*** *With probability* $\geq 1 - \frac{\sqrt{8k}}{\sqrt{\pi}e^{8k}}$*, for all* $t \geq 0$ *there exists* α_i*,* $1 \leq i \leq 4$ *such that* $|\alpha_i| \leq \eta$ *and the following holds:*

1. *For* $i \in \{1, 3\}$ *and* $j \in W_0^+(i)$*, it holds that* $\boldsymbol{w}_t^{(j)} = \boldsymbol{w}_0^{(j)} + a_i(t)\eta\boldsymbol{x}_i + \alpha_i\boldsymbol{x}_2$.
2. *For* $i \in \{2, 4\}$ *and* $j \in U_0^+(i)$*, it holds that* $\boldsymbol{u}_t^{(j)} = \boldsymbol{u}_0^{(j)} + a_i(t)\eta\boldsymbol{x}_i + \alpha_i\boldsymbol{x}_1$.

Proof. By Lemma B.2, with probability $\geq 1 - \frac{\sqrt{8k}}{\sqrt{\pi}e^{8k}}$, for all $1 \leq j \leq k$ and $1 \leq i \leq 4$ it holds that $\left|\boldsymbol{w}_0^{(j)} \cdot \boldsymbol{x}_i\right| \leq \frac{\sqrt{2}\eta}{4}$ and $\left|\boldsymbol{u}_0^{(j)} \cdot \boldsymbol{x}_i\right| \leq \frac{\sqrt{2}\eta}{4}$. It suffices to prove the claim for $W_t^+(1)$. The other cases follow by a symmetry. The proof is by induction. Assume that $j \in W_t^+(1)$. For $t = 0$ the claim holds with $\alpha_1^t = 0$. For a point $(\boldsymbol{x}, y)$ let $\ell_{(\boldsymbol{x},y)} = \max\{1 - yN_W(\boldsymbol{x}), 0\}$. Then it holds that $\frac{\partial \ell_{(\boldsymbol{x},y)}}{\partial \boldsymbol{w}^{(i)}}(W) = -y\sigma'(\boldsymbol{w}^{(i)} \cdot \boldsymbol{x})\boldsymbol{x}\mathbb{1}_{yN_W(\boldsymbol{x})<1}$. Assume without loss of generality that $\alpha_1 > 0$. Define $\beta_1 = \mathbb{1}_{N_W(\boldsymbol{x}_1)<1}$ and $\beta_2 = \mathbb{1}_{N_W(\boldsymbol{x}_2)>-1}$. Using these notations, we have

$$\begin{aligned}\boldsymbol{w}_{t+1}^{(j)} &= \boldsymbol{w}_t^{(j)} + \beta_1\eta\boldsymbol{x}_1 - \beta_2\eta\boldsymbol{x}_2 \\ &= \boldsymbol{w}_0^{(j)} + (a_i(t) + \beta_1)\boldsymbol{x}_i + (\alpha_i - \beta_2\eta)\boldsymbol{x}_2\end{aligned}$$

and for any values of $\beta_1, \beta_2 \in \{0, 1\}$ the induction step follows.

□

For each point $\boldsymbol{x}_i$, define the following sums:

$$S_t^+(i) = \sum_{j \in W_t^+(i)} \sigma\left(\boldsymbol{w}_t^{(j)} \cdot \boldsymbol{x}_i\right)$$

$$R_t^+(i) = \sum_{j \in U_t^+(i)} \sigma\left(\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{x}_i\right)$$

We will prove the following lemma regarding $S_t^+(1), R_t^+(1)$ for $i = 1$. By symmetry, analogous lemmas follow for $i \neq 1$.

Lemma B.4. *The following holds with probability* $\geq 1 - \frac{\sqrt{8k}}{\sqrt{\pi}e^{8k}}$*:*

1. *For all* $t \geq 0$, $R_t^+(1) \leq 2k\eta$.
2. *If* $yN_{W_t}(\boldsymbol{x}_1) < 1$, *then* $S_{t+1}^+(1) \geq S_t^+(1) + \left|W_0^+(1)\right|\eta$. *Otherwise, if* $-yN_{W_t}(\boldsymbol{x}_1) \geq 1$ *then* $S_{t+1}^+(1) = S_t^+(1)$.

Proof. 1. Assume by contradiction that there exists $t > 0$, such that $R_t^+(1) > 2k\eta$. It follows that there exists $j \in U_t^+(1)$ such that $\sigma\left(\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{x}_1\right) > 2\eta$. However, this contradicts Lemma B.3 and Lemma B.2, because with probability 1, there exists $l \in \{2, 4\}$ such that $j \in U_0^+(l)$.

2. This follows by Lemma B.3. We note that by this lemma, if $j \in W_0^+(1)$ then $j \in W_t^+(1)$ for all $t > 0$.

□

Proposition B.5. *Assume that* $k > 16$. *With probability* $\geq 1 - \frac{\sqrt{8k}}{\sqrt{\pi}e^{8k}} - 8e^{-8}$, *for all* i, *if until iteration* T *there were at least* $l \geq \frac{1}{c_\eta}\left(\frac{4\sqrt{k}}{\sqrt{k}-4}\right)$ *iterations, in which* $yN_{W_t}(\boldsymbol{x}_i) < 1$, *then it holds that* $yN_{W_t}(\boldsymbol{x}_i) \geq 1$ *for all* $t \geq T$.

Proof. Without loss of generality assume that $i = 1$. By Lemma B.4 and Lemma E.3, with probability $\geq 1 - \frac{\sqrt{8k}}{\sqrt{\pi}e^{8k}} - 8e^{-8}$, if $yN_{W_t}(\boldsymbol{x}_1) < 1$ then $S_{t+1}^+(1) \geq S_t^+(1) + \left(\frac{k}{2} - 2\sqrt{k}\right)\eta$. Therefore, by Lemma B.4, for all $t \geq T$

$$\begin{aligned} N_{W_t}(\boldsymbol{x}_1) &= S_t^+(1) - R_t^+(1) \\ &\geq \left(\frac{k}{2} - 2\sqrt{k}\right) l\eta - 2k\eta \\ &\geq 1 \end{aligned}$$

where the last ineqaulity follows by the assumption on l. □

Theorem B.6. ***Convergence and clustering. Theorem 3.3 restated.*** *Assume that* $k > 16$. *With probability* $\geq 1 - \frac{\sqrt{8k}}{\sqrt{\pi}e^{8k}} - 8e^{-8}$, *after* $T > \frac{1}{c_\eta}\left(\frac{16\sqrt{k}}{\sqrt{k}-4}\right)$ *iterations, gradient descent converges to a global minimum. Furthermore, for* $i \in \{1, 3\}$ *and all* $j \in W_0^+(i)$, *the angle between* $\boldsymbol{w}_T^{(j)}$ *and* $\boldsymbol{x}_i$ *is at most* $\arccos\left(\frac{1-2c_\eta}{1+c_\eta}\right)$. *Similarly, for* $i \in \{2, 4\}$ *and all* $j \in U_0^+(i)$, *the angle between* $\boldsymbol{u}_T^{(j)}$ *and* $\boldsymbol{x}_i$ *is at most* $\arccos\left(\frac{1-2c_\eta}{1+c_\eta}\right)$.

Proof. Proposition B.5 implies that there are at most $\frac{1}{c_\eta}\left(\frac{16\sqrt{k}}{\sqrt{k}-4}\right)$ iterations in which there exists $(\boldsymbol{x}_i, y_i)$ such that $y_i N_{W_t}(\boldsymbol{x}_i) < 1$. After at most that many iterations, gradient descent converges to a global minimum.

Without loss of generality, we prove the clustering claim for $i = 1$ and all $j \in W_0^+(1)$. At a global minimum, $N_{W_T}(\boldsymbol{x}_1) \geq 1$. Therefore, by Lemma B.3 and Lemma B.4 it follows that

$$2\eta(a_i(T)+1)\left|W_0^+(1)\right| \geq S_t^+(1) \geq 1$$

and thus $a_i(T) \geq \frac{1}{2c_\eta} - 1$. Therefore, for any $j \in W_0^+(1)$, the cosine of the angle between $\boldsymbol{w}_T^{(j)}$ and $\boldsymbol{x}_1$ is at least

$$\frac{(\boldsymbol{w}_0^{(j)} + a_1(T)\eta\boldsymbol{x}_1 + \alpha_1\boldsymbol{x}_2)\cdot\boldsymbol{x}_1}{\sqrt{2}(\|\boldsymbol{w}_0^{(j)}\| + \sqrt{2}a_i(T)\eta + \sqrt{2}\eta)} \geq \frac{2a_1(T)}{2a_1(T)+3} \geq \frac{1-2c_\eta}{1+c_\eta}$$

where we used the triangle inequality, Lemma B.3 and Lemma B.2. The last inequality follows since $f(x) = \frac{2x}{2x+3}$ is monotonically increasing. The claim follows. □

B.2 Small Network

Lemma B.7. ***Non-exploration at initialization.*** *With probability at least* 0.75*, there exists* $i \in \{1,3\}$ *such that* $W_0^+(i) = \emptyset$ *or* $i \in \{2,4\}$ *such that* $U_0^+(i) = \emptyset$*.*

Proof. Since the sign of a one dimensional Gaussian random variable is a Bernoulli random variable, the probability that $W_0^+(i) \neq \emptyset$ for $i \in \{1,3\}$ and $U_0^+(i) \neq \emptyset$ for $i \in \{2,4\}$ is $\frac{1}{4}$. The claim follows. □

Theorem B.8. *Assume that* $k = 2$*. With probability* ≥ 0.75*, gradient descent converges to a local minimum.*

Proof. As in the proof of Theorem 3.3, for $i \in \{1,3\}$ if $W_0^+(i) \neq \emptyset$, then eventually, $y_i N_{W_t}(\boldsymbol{x}_i) \geq 1$. Similarly, for $i \in \{2,4\}$ if $U_0^+(i) \neq \emptyset$, then eventually, $y_i N_{W_t}(\boldsymbol{x}_i) \geq 1$. However, if without loss of generality $W_0^+(1) = \emptyset$, then for all t,

$$N_{W_t}(\boldsymbol{x}_1) = S_t^+(1) - R_t^+(1) \leq 0$$

Furthermore, there exists the first iteration t' such that $y_i N_{W_{t'}}(\boldsymbol{x}_i) \geq 1$ for $i = 3$ (since $W_0^+(3) \neq \emptyset$) and any $i \in \{2,4\}$ such that $U_0^+(i) \neq \emptyset$. Then, in iteration $t'+1$ for all $1 \leq j \leq 2$ it holds that $\boldsymbol{u}_{t'+1}^{(j)}\boldsymbol{x}_i < 0$ and $\boldsymbol{w}_{t'+1}^{(j)}\boldsymbol{x}_i < 0$ for $i = 1$ and $i \in \{2,4\}$ such that $U_0^+(i) = \emptyset$ (here we use the fact that $\left|\boldsymbol{u}_t^{(j)}\boldsymbol{x}_1\right| \leq \eta$ for all t by Lemma B.2. Similarly, for $\left|\boldsymbol{w}_t^{(j)}\boldsymbol{x}_i\right|$ where $i \in \{2,4\}$). Therefore at $t'+1$ we are at a local minimum. □

C Proofs and Experiments for Section 4

C.1 VC Dimension

As noted in Remark 4.1, the VC dimension of the model we consider is at most 15. To see this, we first define for any $\boldsymbol{z} \in \{\pm 1\}^{2d}$ the set $P_{\boldsymbol{z}} \subseteq \{\pm 1\}^2$ which contains all the distinct two dimensional binary patterns that $\boldsymbol{z}$ has. For example, for a positive diverse point $\boldsymbol{z}$ it holds that $P_{\boldsymbol{z}} = \{\pm 1\}^2$. Now, for any points $\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)} \in \{\pm 1\}^{2d}$ such that $P_{\boldsymbol{z}^{(1)}} = P_{\boldsymbol{z}^{(2)}}$ and for any filter $\boldsymbol{w} \in \mathbb{R}^2$ it holds that $\max_j \sigma\left(\boldsymbol{w}\cdot\boldsymbol{z}_j^{(1)}\right) = \max_j \sigma\left(\boldsymbol{w}\cdot\boldsymbol{z}_j^{(2)}\right)$. Therefore, for any W, $N_W(\boldsymbol{z}^{(1)}) = N_W(\boldsymbol{z}^{(2)})$. Specifically, this implies that if both $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$ are diverse then $N_W(\boldsymbol{z}^{(1)}) = N_W(\boldsymbol{z}^{(2)})$. Since there are 15 non-empty subsets of $\{\pm 1\}^2$, it follows that for any k the network can shatter a set of at most 15 points, or equivalently, its VC dimension is at most 15. Despite these expressive power limitations, there is a generalization gap between small and large networks in this setting, as can be seen in Figure 1.

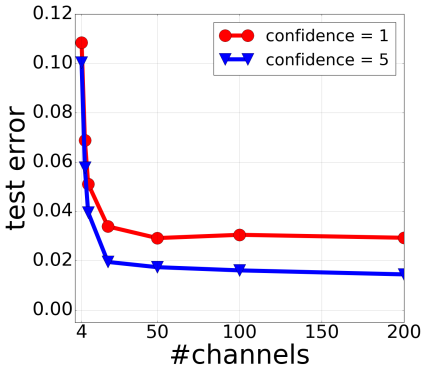


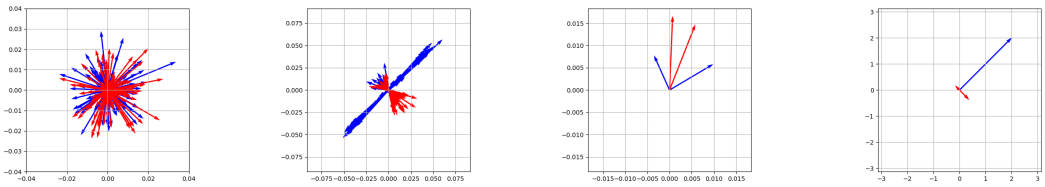
Figure 1: Higher confidence of hinge-loss results in better performance in the XORD problem.

C.2 Hinge Loss Confidence

Figure 1 shows that setting $\gamma = 5$ gives better performance than setting $\gamma = 1$ in the XORD problem. The setting is similar to the setting of Section A. Each point is an average test error of 100 runs.

D Experiments for Section 5

Here we show an example of a training set that has a non-diverse negative point. The training set contains 6 diverse positive points, 5 diverse negative points and a negative non-diverse point that only contains the pattern $\boldsymbol{p}_4$. We implemented the setting of Section 4 and ran gradient descent on this training set. In Figure 2 we show the results. The large network recovers f^*, while the small does not. This is despite the fact that both networks achieve zero training error.



(a) (b) (c) (d)

Figure 2: Overparameterization and generalization in XORD problem. The vectors in blue are the vectors $\boldsymbol{w}_t^{(i)}$ and in red are the vectors $\boldsymbol{u}_t^{(i)}$. (a) Exploration at initialization (t=0) for $k = 100$ (b) Clustering and convergence to global minimum that recovers f^* for $k = 100$ (c) Non-sufficient exploration at initialization (t=0) for $k = 2$. (d) Convergence to global minimum with non-zero test error for $k = 2$.

E Proof of Theorem 6.3

We first restate the theorem.

Theorem E.1. ***(Theorem 6.3 restated and extended.)*** *With probability at least* $\left(1 - c - 16e^{-8}\right)$ *after running gradient descent for* $T \geq \frac{28(\gamma+1+8c_\eta)}{c_\eta}$ *iterations, it converges to a global minimum which*

satisfies $\operatorname{sign}\left(N_{W_T}(\boldsymbol{x})\right) = f^*(\boldsymbol{x})$ *for all* $\boldsymbol{x} \in \{\pm 1\}^{2d}$. *Furthermore, for* $i \in \{1,3\}$ *and all* $j \in W_0^+(i)$, *the angle between* $\boldsymbol{w}_T^{(j)}$ *and* $\boldsymbol{p}_i$ *is at most* $\arccos\left(\frac{\gamma-1-2c_\eta}{\gamma-1+c_\eta}\right)$.

We will first need a few notations. Define $\boldsymbol{p}_1 = (1,1), \boldsymbol{x}_2 = (1,-1), \boldsymbol{p}_3 = (-1,-1), \boldsymbol{p}_4 = (-1,1)$ and the following sets:

$$W_t^+(i) = \left\{ j \mid \arg\max_{1 \le l \le 4} \boldsymbol{w}_t^{(j)} \cdot \boldsymbol{p}_l = i \right\},\ U_t^+(i) = \left\{ j \mid \arg\max_{1 \le l \le 4} \boldsymbol{u}_t^{(j)} \cdot \boldsymbol{p}_l = i \right\}$$

$$W_t^-(i) = \left\{ j \mid \arg\max_{l \in \{2,4\}} \boldsymbol{w}_t^{(j)} \cdot \boldsymbol{p}_l = i \right\},\ U_t^-(i) = \left\{ j \mid \arg\max_{l \in \{2,4\}} \boldsymbol{u}_t^{(j)} \cdot \boldsymbol{p}_l = i \right\}$$

We can use these definitions to express more easily the gradient updates. Concretely, let $j \in W_t^+(i_1) \cap W_t^-(i_2)$ then the gradient update is given as follows:[1]

$$\boldsymbol{w}_{t+1}^{(j)} = \boldsymbol{w}_t^{(j)} + \eta \boldsymbol{p}_{i_1} \mathbb{1}_{N_W(\boldsymbol{x}^+) < \gamma} - \eta \boldsymbol{p}_{i_2} \mathbb{1}_{N_W(\boldsymbol{x}^-) < 1} \tag{2}$$

Similarly, for $j \in U_t^+(i_1) \cap U_t^-(i_2)$ the gradient update is given by:

$$\boldsymbol{u}_{t+1}^{(j)} = \boldsymbol{u}_t^{(j)} - \eta \boldsymbol{p}_{i_1} \mathbb{1}_{N_W(\boldsymbol{x}^+) < \gamma} + \eta \boldsymbol{p}_{i_2} \mathbb{1}_{N_W(\boldsymbol{x}^-) < 1} \tag{3}$$

We denote by $\boldsymbol{x}^+$ a positive diverse point and $\boldsymbol{x}^-$ a negative diverse point. Define the following sums for $\phi \in \{+,-\}$:

$$S_t^\phi = \sum_{j \in W_t^+(1) \cup W_t^+(3)} \left[\max\left\{ \sigma\left(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}_1^\phi\right), ..., \sigma\left(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}_d^\phi\right) \right\} \right]$$

$$P_t^\phi = \sum_{j \in U_t^+(1) \cup U_t^+(3)} \left[\max\left\{ \sigma\left(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}_1^\phi\right), ..., \sigma\left(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}_d^\phi\right) \right\} \right]$$

$$R_t^\phi = \sum_{j \in W_t^+(2) \cup W_t^+(4)} \left[\max\left\{ \sigma\left(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}_1^\phi\right), ..., \sigma\left(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}_d^\phi\right) \right\} \right] - \sum_{j \in U_t^+(2) \cup U_t^+(4)} \left[\max\left\{ \sigma\left(\boldsymbol{u}^{(i)} \cdot \boldsymbol{x}_1^\phi\right), ..., \sigma\left(\boldsymbol{u}^{(i)} \cdot \boldsymbol{x}_d^\phi\right) \right\} \right]$$

Note that $R_t^+ = R_t^-$ since for $\boldsymbol{z} \in \{\boldsymbol{x}^+, \boldsymbol{x}^-\}$ there exists i_1, i_2 such that $\boldsymbol{z}_{i_1} = \boldsymbol{p}_2$, $\boldsymbol{z}_{i_2} = \boldsymbol{p}_4$.

Without loss of generality, we can assume that the training set consists of one positive diverse point $\boldsymbol{x}^+$ and one negative diverse point $\boldsymbol{x}^-$. This follows since the network and its gradient have the same value for two different positive diverse points and two different negative points. Therefore, this holds for the loss function defined in Eq. 4 as well.

We let $a^+(t)$ be the number of iterations $0 \le t' \le t$ such that $N_{W_{t'}}(\boldsymbol{x}^+) < \gamma$.

We will now proceed to prove the theorem. In Section E.0.1 we prove results on the filters at initialization. In Section E.0.2 we prove several lemmas that exhibit the clustering dynamics. In Section E.0.3 we prove upper bounds on S_t^-, P_t^+ and P_t^- for all iterations t. In Section E.0.4 we characterize the dynamics of S_t^+ and in Section E.0.5 we prove an upper bound on it together with upper bounds on $N_{W_t}(\boldsymbol{x}^+)$ and $-N_{W_t}(\boldsymbol{x}^-)$ for all iterations t.

We provide an optimization guarantee for gradient descent in Section E.0.6. We prove generalization guarantees for the points in the positive class and negative class in Section E.0.7 and Section E.0.8, respectively. We complete the proof of the theorem in Section E.0.9 with proofs for the clustering effect at the global minimum.

[1] Note that with probability 1, $\sigma'(\boldsymbol{w}_t^{(j)} \cdot \boldsymbol{p}_{i_1}) = 1$, $\sigma'(\boldsymbol{w}_t^{(j)} \cdot \boldsymbol{p}_{i_2}) = 1$ for all t, and therefore we omit these from the gradient update. This follows since $\sigma'(\boldsymbol{w}_t^{(j)} \cdot \boldsymbol{p}_{i_1}) = 0$ for some t if and only if $\boldsymbol{w}_0^{(j)} \cdot \boldsymbol{p}_{i_1}$ is an integer multiple of η.

E.0.1 Initialization Guarantees

Lemma E.2. ***Exploration. Lemma 6.1 restated and extended.*** *With probability at least* $1 - 4e^{-8}$*, it holds that*

$$\left| \left| W_0^+(1) \cup W_0^+(3) \right| - \frac{k}{2} \right| \leq 2\sqrt{k}$$

and

$$\left| \left| U_0^+(1) \cup U_0^+(3) \right| - \frac{k}{2} \right| \leq 2\sqrt{k}$$

Proof. Without loss of generality consider $\left| W_0^+(1) \cup W_0^+(3) \right|$. Since $\mathbb{P}\left[j \in W_0^+(1) \cup W_0^+(3) \right] = \frac{1}{2}$, we get by Hoeffding's inequality

$$\mathbb{P}\left[\left| \left| W_0^+(1) \cup W_0^+(3) \right| - \frac{k}{2} \right| < 2\sqrt{k} \right] \leq 2e^{-\frac{2(2^2 k)}{k}} = 2e^{-8}$$

The result now follows by the union bound. □

Lemma E.3. *With probability* $\geq 1 - \frac{\sqrt{8k}}{\sqrt{\pi} e^{8k}}$*, for all* $1 \leq j \leq k$ *and* $1 \leq i \leq 4$ *it holds that* $\left| \boldsymbol{w}_0^{(j)} \cdot \boldsymbol{p}_i \right| \leq \frac{\sqrt{2}\eta}{4}$ *and* $\left| \boldsymbol{u}_0^{(j)} \cdot \boldsymbol{p}_i \right| \leq \frac{\sqrt{2}\eta}{4}$.

Proof. Let Z be a random variable distributed as $\mathcal{N}(0, \sigma^2)$. Then by Proposition 2.1.2 in Vershynin (2017), we have

$$\mathbb{P}\left[|Z| \geq t \right] \leq \frac{2\sigma}{\sqrt{2\pi} t} e^{-\frac{t^2}{2\sigma^2}}$$

Therefore, for all $1 \leq j \leq k$ and $1 \leq i \leq 4$,

$$\mathbb{P}\left[\left| \boldsymbol{w}_0^{(j)} \cdot \boldsymbol{p}_i \right| \geq \frac{\sqrt{2}\eta}{4} \right] \leq \frac{1}{\sqrt{8\pi k}} e^{-8k}$$

and

$$\mathbb{P}\left[\left| \boldsymbol{u}_0^{(j)} \cdot \boldsymbol{p}_i \right| \geq \frac{\sqrt{2}\eta}{4} \right] \leq \frac{1}{\sqrt{8\pi k}} e^{-8k}$$

The result follows by applying a union bound over all $2k$ weight vectors and the four points $\boldsymbol{p}_i$, $1 \leq i \leq 4$. □

From now on we assume that the highly probable event in Lemma E.3 holds.

Lemma E.4. $N_{W_t}(\boldsymbol{x}^+) < 1$ *and* $-N_{W_t}(\boldsymbol{x}^-) < 1$ *for* $0 \leq t \leq 2$.

Proof. By Lemma E.3 we have

$$\begin{aligned} N_{W_0}(\boldsymbol{x}^+) &= \sum_{i=1}^{k} \left[\max\left\{ \sigma\left(\boldsymbol{w}_0^{(i)} \cdot \boldsymbol{x}_1^+ \right), ..., \sigma\left(\boldsymbol{w}_0^{(i)} \cdot \boldsymbol{x}_d^+ \right) \right\} - \max\left\{ \sigma\left(\boldsymbol{u}_0^{(i)} \cdot \boldsymbol{x}_1^+ \right), ..., \sigma\left(\boldsymbol{u}_0^{(i)} \cdot \boldsymbol{x}_d^+ \right) \right\} \right] \\ &\leq \frac{\sqrt{2}\eta k}{4} < \gamma \end{aligned}$$

and similarly $-N_{W_0}(\boldsymbol{x}^-) < 1$. Therefore, by Eq. 2 and Eq. 3 we get:

1. For $i \in \{1, 3\}$, $l \in \{2, 4\}$, $j \in W_0^+(i) \cap W_0^-(l)$, it holds that $\boldsymbol{w}_1^{(j)} = \boldsymbol{w}_0^{(j)} - \eta \boldsymbol{p}_l + \eta \boldsymbol{p}_i$.
2. For $i \in \{2, 4\}$ and $j \in W_0^+(i)$, it holds that $\boldsymbol{w}_1^{(j)} = \boldsymbol{w}_0^{(j)}$.

3. For $i \in \{1,3\}$, $l \in \{2,4\}$, $j \in U_0^+(i) \cap U_0^-(l)$, it holds that $\boldsymbol{u}_1^{(j)} = \boldsymbol{u}_0^{(j)} - \eta \boldsymbol{p}_i + \eta \boldsymbol{p}_l$.
4. For $i \in \{2,4\}$ and $j \in U_0^+(i)$, it holds that $\boldsymbol{u}_2^{(j)} = \boldsymbol{u}_0^{(j)}$.

Applying Lemma E.3 again and using the fact that $\eta \leq \frac{1}{8k}$ we have $N_{W_1}(\boldsymbol{x}^+) < \gamma$ and $-N_{W_1}(\boldsymbol{x}^-) < 1$. Therefore we get,

1. For $i \in \{1,3\}$, $l \in \{2,4\}$, $j \in W_0^+(i) \cap W_0^-(l)$, it holds that $\boldsymbol{w}_2^{(j)} = \boldsymbol{w}_0^{(j)} + 2\eta \boldsymbol{p}_i$.
2. For $i \in \{2,4\}$ and $j \in W_0^+(i)$, it holds that $\boldsymbol{w}_2^{(j)} = \boldsymbol{w}_0^{(j)}$.
3. For $i \in \{1,3\}$, $l \in \{2,4\}$, $j \in U_0^+(i) \cap U_0^-(l)$, it holds that $\boldsymbol{u}_2^{(j)} = \boldsymbol{u}_0^{(j)} - \eta \boldsymbol{p}_i + \eta \boldsymbol{p}_l$.
4. For $i \in \{2,4\}$ and $j \in U_0^+(i)$, it holds that $\boldsymbol{u}_2^{(j)} = \boldsymbol{u}_0^{(j)}$.

As before, by Lemma E.3 we have $N_{W_2}(\boldsymbol{x}^+) < \gamma$ and $-N_{W_2}(\boldsymbol{x}^-) < 1$. □

E.0.2 Clustering Dynamics Lemmas

In the following lemmas we assume that the highly probable event in Lemma E.3 holds. We therefore do not mention the probability in the statements of the lemmas.

Lemma E.5. ***Clusetering. Lemma 6.2 restated and extended.*** *For all $t \geq 0$ there exists $\alpha_i^{(t)}$, $i \in \{1,3\}$ such that $\left|\alpha_i^{(t)}\right| \leq \eta$ and the following holds:*

1. *For $i \in \{1,3\}$ and $j \in W_0^+(i)$, it holds that $\boldsymbol{w}_t^{(j)} = \boldsymbol{w}_0^{(j)} + a^+(t)\eta \boldsymbol{p}_i + \alpha_i^{(t)} \boldsymbol{p}_2$.*
2. *For $i \in \{2,4\}$ and $j \in W_0^+(i)$, it holds that $\boldsymbol{w}_t^{(j)} = \boldsymbol{w}_0^{(j)} + m\boldsymbol{p}_2$ for $m \in \mathbb{Z}$.*
3. *$W_t^+(i) = W_0^+(i)$ for $i \in \{1,3\}$.*

Proof. By Lemma E.3, with probability $\geq 1 - \frac{\sqrt{8k}}{\sqrt{\pi} e^{8k}}$, for all $1 \leq j \leq k$ and $1 \leq i \leq 4$ it holds that $\left|\boldsymbol{w}_0^{(j)} \cdot \boldsymbol{x}_i\right| \leq \frac{\sqrt{2}\eta}{4}$ and $\left|\boldsymbol{u}_0^{(j)} \cdot \boldsymbol{x}_i\right| \leq \frac{\sqrt{2}\eta}{4}$. We will first prove the first claim and that $W_0^+(i) \subseteq W_t^+(i)$ for all $t \geq 1$. To prove this, we will show by induction on $t \geq 1$, that for all $j \in W_0^+(i) \cap W_0^+(l)$, where $l \in \{2,4\}$ the following holds:

1. $j \in W_t^+(i)$.
2. $\boldsymbol{w}_t^{(j)} \cdot \boldsymbol{p}_l = \boldsymbol{w}_0^{(j)} \cdot \boldsymbol{p}_l - 2\eta$ or $\boldsymbol{w}_t^{(j)} \cdot \boldsymbol{p}_l = \boldsymbol{w}_t^{(0)} \cdot \boldsymbol{p}_l$.
3. $\boldsymbol{w}_t^{(j)} = \boldsymbol{w}_0^{(j)} + a^+(t)\eta \boldsymbol{p}_i + \alpha_i^{(t)} \boldsymbol{p}_2$, where $\left|\alpha_i^{(t)}\right| \leq \eta$.
4. $\boldsymbol{w}_t^{(j)} \cdot \boldsymbol{p}_i > 2\eta$.

The claim holds for $t = 1$ by the proof of Lemma E.4. Assume it holds for $t = T$. By the induction hypothesis there exists an $l' \in \{2,4\}$ such that $j \in W_T^+(i) \cap W_T^-(l')$. By Eq. 2 we have,

$$\boldsymbol{w}_{T+1}^{(j)} = \boldsymbol{w}_T^{(j)} + a\eta \boldsymbol{p}_i + b\eta \boldsymbol{p}_{l'} \tag{4}$$

where $a = a^+(t+1) - a^+(t)$ and $b \in \{-1, 0\}$. From this follows the third claim of the induction proof and the first claim of the lemma.

If $\boldsymbol{w}_T^{(j)} \cdot \boldsymbol{p}_l = \boldsymbol{w}_0^{(j)} \cdot \boldsymbol{p}_l$ then $l' = l$ and either $\boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{p}_l = \boldsymbol{w}_0^{(j)} \cdot \boldsymbol{p}_l$ if $b = 0$ or $\boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{p}_l = \boldsymbol{w}_0^{(j)} \cdot \boldsymbol{p}_l - 2\eta$ if $b = -1$. Otherwise, assume that $\boldsymbol{w}_T^{(j)} \cdot \boldsymbol{p}_l = \boldsymbol{w}_0^{(j)} \cdot \boldsymbol{p}_l - 2\eta$. By Lemma E.3 we have $0 < \boldsymbol{w}_0^{(j)} \cdot \boldsymbol{p}_l < \frac{\sqrt{2}\eta}{4}$. Therefore $-2\eta < \boldsymbol{w}_T^{(j)} \cdot \boldsymbol{p}_l < 0$ and $l' \neq l$. It follows that either $\boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{p}_l = \boldsymbol{w}_0^{(j)} \cdot \boldsymbol{p}_l - 2\eta$ if $b = 0$

or $\boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{p}_l = \boldsymbol{w}_0^{(j)} \cdot \boldsymbol{p}_l$ if $b = -1$. In both cases, we have $\left|\boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{p}_l\right| < 2\eta$. Furthermore, by Eq. 4, $\boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{p}_i \geq \boldsymbol{w}_T^{(j)} \cdot \boldsymbol{p}_i > 2\eta$. Hence, $\arg\max_{1 \leq l \leq 4} \boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{p}_l = i$ which by definition implies that $j \in W_{T+1}^+(i)$. This concludes the proof by induction which shows that $W_0^+(i) \subseteq W_t^+(i)$ for all $t \geq 1$.

To prove that $W_t^+(i) = W_0^+(i)$ for $i \in \{1, 3\}$, it suffices to show that $W_0^+(2) \cup W_0^+(4) \subseteq W_t^+(2) \cup W_t^+(4)$. This follows since $\bigcup_{i=1}^4 W_t^+(i) = \{1, 2, ..., k\}$. We will show by induction on $t \geq 1$, that for all $j \in W_0^+(2) \cup W_0^+(4)$, the following holds:

1. $j \in W_t^+(2) \cap W_t^+(4)$.
2. $\boldsymbol{w}_t^{(j)} = \boldsymbol{w}_0^{(j)} + m\boldsymbol{p}_2$ for $m \in \mathbb{Z}$.

This will conclude the proof of the lemma. The claim holds for $t = 1$ by the proof of Lemma E.4. Assume it holds for $t = T$. By the induction hypothesis $j \in W_T^+(2) \cap W_T^+(4)$. Assume without loss of generality that $j \in W_T^+(2)$. This implies that $j \in W_T^-(2)$ as well. Therefore, by Eq. 2 we have

$$\boldsymbol{w}_{T+1}^{(j)} = \boldsymbol{w}_T^{(j)} + a\eta\boldsymbol{p}_2 + b\eta\boldsymbol{p}_2 \tag{5}$$

where $a \in \{0, 1\}$ and $b \in \{0, -1\}$. By the induction hypothesis, $\boldsymbol{w}_{T+1}^{(j)} = \boldsymbol{w}_0^{(j)} + m\boldsymbol{p}_2$ for $m \in \mathbb{Z}$. If $a = 1$ or $b = 0$ we have for $i \in \{1, 3\}$,

$$\boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{p}_2 \geq \boldsymbol{w}_T^{(j)} \cdot \boldsymbol{p}_2 > \boldsymbol{w}_T^{(j)} \cdot \boldsymbol{p}_i = \boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{p}_i$$

where the second inequality follows by Eq. 5 (we note that this inequality is strict with probability 1). This implies that $j \in W_{T+1}^+(2) \cap W_{T+1}^+(4)$.

Otherwise, assume that $a = 0$ and $b = -1$. By Lemma E.3 we have $\boldsymbol{w}_0^{(j)} \cdot \boldsymbol{p}_2 < \frac{\sqrt{2}\eta}{4}$. Since $j \in W_T^+(2)$, it follows by the induction hypothesis that $\boldsymbol{w}_T^{(j)} = \boldsymbol{w}_0^{(j)} + m\boldsymbol{p}_2$, where $m \in \mathbb{Z}$ and $m \geq 0$. To see this, note that if $m < 0$, then $\boldsymbol{w}_T^{(j)} \cdot \boldsymbol{p}_2 < 0$ and $j \notin W_T^+(2)$, which is a contradiction. Let $i \in \{1, 3\}$. If $m = 0$, then $\boldsymbol{w}_{T+1}^{(j)} = \boldsymbol{w}_0^{(j)} - \boldsymbol{p}_2$, $\boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{p}_4 > \frac{\sqrt{2}\eta}{4}$ and $\boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{p}_i = \boldsymbol{w}_0^{(j)} \cdot \boldsymbol{p}_i < \frac{\sqrt{2}\eta}{4}$ by Lemma E.3. Therefore, $j \in W_{T+1}^+(4)$.

Otherwise, if $m > 0$, then $\boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{p}_2 \geq \boldsymbol{w}_0^{(j)} \cdot \boldsymbol{p}_2 > \boldsymbol{w}_0^{(j)} \cdot \boldsymbol{p}_i = \boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{p}_i$. Hence, $j \in W_{T+1}^+(2)$, which concludes the proof. □

Lemma E.6. *For all $t \geq 0$ we have*

1. $\boldsymbol{u}_t^{(j)} = \boldsymbol{u}_0^{(j)} + m\eta\boldsymbol{p}_2$ *for* $m \in \mathbb{Z}$.
2. $U_0^+(2) \cup U_0^+(4) \subseteq U_t^+(2) \cup U_t^+(4)$.

.

Proof. Let $j \in U_0^+(2) \cup U_0^+(4)$. It suffices to prove that $\boldsymbol{u}_t^{(j)} = \boldsymbol{u}_0^{(j)} + \alpha_t\eta\boldsymbol{p}_2$ for $\alpha_t \in \mathbb{Z}$. This follows since the inequalities $\left|\boldsymbol{u}_0^{(j)} \cdot \boldsymbol{p}_1\right| < \left|\boldsymbol{u}_0^{(j)} \cdot \boldsymbol{p}_2\right| \leq \frac{\sqrt{2}\eta}{4}$ imply that in this case $j \in U_t^+(2) \cup U_t^+(4)$. Assume by contradiction that there exist an iteration t for which $\boldsymbol{u}_t^{(j)} = \boldsymbol{u}_0^{(j)} + \alpha_t\eta\boldsymbol{p}_2 + \beta_t\eta\boldsymbol{p}_i$ where $\beta_t \in \{-1, 1\}$, $\alpha_t \in \mathbb{Z}$, $i \in \{1, 3\}$ and $\boldsymbol{u}_{t-1}^{(j)} = \boldsymbol{u}_0^{(j)} + \alpha_{t-1}\eta\boldsymbol{p}_2$ where $\alpha_{t-1} \in \mathbb{Z}$. [2] Since the coefficient of $\boldsymbol{p}_i$ changed in iteration t, we have $j \in U_{t-1}^+(1) \cup U_{t-1}^+(3)$. However, this contradicts the claim above which shows that if $\boldsymbol{u}_{t-1}^{(j)} = \boldsymbol{u}_0^{(j)} + \alpha_{t-1}\eta\boldsymbol{p}_2$, then $j \in U_{t-1}^+(2) \cup U_{t-1}^+(4)$. □

Lemma E.7. *Let $i \in \{1, 3\}$ and $l \in \{2, 4\}$. For all $t \geq 0$, if $j \in U_0^+(i) \cap U_0^-(l)$, then there exists $a_t \in \{0, -1\}$, $b_t \in \mathbb{N}$ such that $\boldsymbol{u}_t^{(j)} = \boldsymbol{u}_0^{(j)} + a_t\eta\boldsymbol{p}_i + b_t\eta\boldsymbol{p}_l$.*

[2] Note that in each iteration β_t changes by at most η.

Proof. First note that by Eq. 3 we generally have $\boldsymbol{u}_t^{(j)} = \boldsymbol{u}_0^{(j)} + \alpha\eta\boldsymbol{p}_i + \beta\eta\boldsymbol{p}_l$ where $\alpha, \beta \in \mathbb{Z}$. Since $\left|\boldsymbol{u}_0^{(j)} \cdot \boldsymbol{p}_1\right| \leq \frac{\sqrt{2}\eta}{4}$, by the gradient update in Eq. 3 it holds that $a_t \in \{0, -1\}$. Indeed, $a_0 = 0$ and by the gradient update if $a_{t-1} = 0$ or $a_{t-1} = -1$ then $a_t \in \{-1, 0\}$.

Assume by contradiction that there exists an iteration $t > 0$ such that $b_t = -1$ and $b_{t-1} = 0$. Note that by Eq. 3 this can only occur if $j \in U_{t-1}^+(l)$. We have $\boldsymbol{u}_{t-1}^{(j)} = \boldsymbol{u}_0^{(j)} + a_{t-1}\eta\boldsymbol{p}_i$ where $a_{t-1} \in \{0, -1\}$. Observe that $\left|\boldsymbol{u}_{t-1}^{(j)} \cdot \boldsymbol{p}_i\right| \geq \left|\boldsymbol{u}_0^{(j)} \cdot \boldsymbol{p}_i\right|$ by the fact that $\left|\boldsymbol{u}_0^{(j)} \cdot \boldsymbol{p}_i\right| \leq \frac{\sqrt{2}\eta}{4}$. Since $\boldsymbol{u}_0^{(j)} \cdot \boldsymbol{p}_i > \boldsymbol{u}_0^{(j)} \cdot \boldsymbol{p}_l = \boldsymbol{u}_{t-1}^{(j)} \cdot \boldsymbol{p}_l$ we have $j \in U_{t-1}^+(1) \cup U_{t-1}^+(3)$, a contradiction. □

E.0.3 Bounding P_t^+, P_t^- and S_t^-

Lemma E.8. *The following holds*

1. $S_t^- \leq \left|W_t^+(1) \cup W_t^+(3)\right| \eta$ *for all* $t \geq 1$.
2. $P_t^+ \leq \left|U_t^+(1) \cup U_t^+(3)\right| \eta$ *for all* $t \geq 1$.
3. $P_t^- \leq \left|U_t^+(1) \cup U_t^+(3)\right| \eta$ *for all* $t \geq 1$.

Proof. In Lemma E.5 we showed that for all $t \geq 0$ and $j \in W_t^+(1) \cup W_t^+(3)$ it holds that $\left|\boldsymbol{w}_t^{(j)} \cdot \boldsymbol{p}_2\right| \leq \eta$. This proves the first claim. The second claim follows similarly. Without loss of generality, let $j \in U_t^+(1)$. By Lemma E.6 it holds that $U_{t'}^+(1) \subseteq U_0^+(1) \cup U_0^+(3)$ for all $t' \leq t$. Therefore, by Lemma E.7 we have $\left|\boldsymbol{u}_t^{(j)}\boldsymbol{p}_1\right| < \eta$, from which the claim follows.

For the third claim, without loss of generality, assume by contradiction that for $j \in U_t^+(1)$ it holds that $\left|\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{p}_2\right| > \eta$. Since $\left|\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{p}_1\right| < \eta$ by Lemma E.7, it follows that $j \in U_t^+(2) \cup U_t^+(4)$, a contradiction. Therefore, $\left|\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{p}_2\right| \leq \eta$ for all $j \in U_t^+(1) \cup U_t^+(3)$, from which the claim follows. □

E.0.4 Dynamics of S_t^+

Lemma E.9. *Let*

$$X_t^+ = \sum_{j \in W_t^+(1)} \left[\max\left\{\sigma\left(\boldsymbol{w}^{(i)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{w}^{(i)} \cdot \boldsymbol{x}_d^+\right)\right\}\right]$$

and

$$Y_t^+ = \sum_{j \in W_t^+(3)} \left[\max\left\{\sigma\left(\boldsymbol{w}^{(i)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{w}^{(i)} \cdot \boldsymbol{x}_d^+\right)\right\}\right]$$

Then for all t, $\frac{X_t^+ - X_0^+}{\left|W_t^+(1)\right|} = \frac{Y_t^+ - Y_0^+}{\left|W_t^+(3)\right|}$.

Proof. We will prove the claim by induction on t. For $t = 0$ this clearly holds. Assume it holds for $t = T$. Let $j_1 \in W_T^+(1)$ and $j_2 \in W_T^+(3)$. By Eq. 2, the gradient updates of the corresponding weight vector are given as follows:

$$\boldsymbol{w}_{T+1}^{(j_1)} = \boldsymbol{w}_T^{(j_1)} + a\eta\boldsymbol{p}_1 + b_1\eta\boldsymbol{p}_2$$

and

$$\boldsymbol{w}_{T+1}^{(j_2)} = \boldsymbol{w}_T^{(j_2)} + a\eta\boldsymbol{p}_3 + b_2\eta\boldsymbol{p}_2$$

where $a \in \{0, 1\}$ and $b_1, b_2 \in \{-1, 0, 1\}$. By Lemma E.5, $j_1 \in W_{T+1}^+(1)$ and $j_2 \in W_{T+1}^+(3)$. Therefore,

$$\max\left\{\sigma\left(\boldsymbol{w}_{T+1}^{(j_1)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{w}_{T+1}^{(j_1)} \cdot \boldsymbol{x}_d^+\right)\right\} = \max\left\{\sigma\left(\boldsymbol{w}_T^{(j_1)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{w}_T^{(j_1)} \cdot \boldsymbol{x}_d^+\right)\right\} + a\eta$$

and

$$\max\left\{\sigma\left(\boldsymbol{w}_{T+1}^{(j_2)}\cdot\boldsymbol{x}_1^+\right),...,\sigma\left(\boldsymbol{w}_{T+1}^{(j_2)}\cdot\boldsymbol{x}_d^+\right)\right\}=\max\left\{\sigma\left(\boldsymbol{w}_{T}^{(j_2)}\cdot\boldsymbol{x}_1^+\right),...,\sigma\left(\boldsymbol{w}_{T}^{(j_2)}\cdot\boldsymbol{x}_d^+\right)\right\}+a\eta$$

By Lemma $E.5$ we have $\left|W_t^+(1)\right|=\left|W_0^+(1)\right|$ and $\left|W_t^+(3)\right|=\left|W_0^+(3)\right|$ for all t. It follows that

$$\begin{aligned}\frac{X_{T+1}^+-X_0^+}{\left|W_{T+1}^+(1)\right|}&=\frac{a\eta\left|W_0^+(1)\right|+X_T^+-X_0^+}{\left|W_0^+(1)\right|}\\&=a\eta+\frac{Y_T^+-Y_0^+}{\left|W_0^+(3)\right|}\\&=\frac{a\eta\left|W_0^+(3)\right|+Y_T^+-Y_0^+}{\left|W_0^+(3)\right|}\\&=\frac{Y_{T+1}^+-Y_0^+}{\left|W_{T+1}^+(3)\right|}\end{aligned}$$

where the second equality follows by the induction hypothesis. This proves the claim. □

Lemma E.10. *The following holds:*

1. *If $N_{W_t}(\boldsymbol{x}^+)<\gamma$ and $-N_{W_t}(\boldsymbol{x}^-)<1$, then $S_{t+1}^+=S_t^++\eta\left|W_t^+(1)\cup W_t^+(3)\right|$.*
2. *If $N_{W_t}(\boldsymbol{x}^+)\geq\gamma$ and $-N_{W_t}(\boldsymbol{x}^-)<1$, then $S_{t+1}^+=S_t^+$.*
3. *If $N_{W_t}(\boldsymbol{x}^+)<\gamma$ and $-N_{W_t}(\boldsymbol{x}^-)\geq1$, then $S_{t+1}^+=S_t^++\eta\left|W_t^+(1)\cup W_t^+(3)\right|$.*

Proof.

1. The equality follows since for each $i\in\{1,3\}$, $l\in\{2,4\}$ and $j\in W_t^+(i)\cap W_t^-(l)$ we have $\boldsymbol{w}_{t+1}^{(j)}=\boldsymbol{w}_t^{(j)}+\eta\boldsymbol{p}_i-\eta\boldsymbol{p}_l$ and $W_{t+1}^+(1)\cup W_{t+1}^+(3)=W_t^+(1)\cup W_t^+(3)$ by Lemma E.5.
2. In this case for each $i\in\{1,3\}$, $l\in\{2,4\}$ and $j\in W_t^+(i)\cap W_t^-(l)$ we have $\boldsymbol{w}_{t+1}^{(j)}=\boldsymbol{w}_t^{(j)}-\eta\boldsymbol{p}_l$ and $W_{t+1}^+(1)\cup W_{t+1}^+(3)=W_t^+(1)\cup W_t^+(3)$ by Lemma E.5.
3. This equality follows since for each $i\in\{1,3\}$, $l\in\{2,4\}$ and $j\in W_t^+(i)\cap W_t^-(l)$ we have $\boldsymbol{w}_{t+1}^{(j)}=\boldsymbol{w}_t^{(j)}+\eta\boldsymbol{p}_i$ and $W_{t+1}^+(1)\cup W_{t+1}^+(3)=W_t^+(1)\cup W_t^+(3)$ by Lemma E.5.

□

E.0.5 Upper Bounds on $N_{W_t}(\boldsymbol{x}^+)$, $-N_{W_t}(\boldsymbol{x}^-)$ and S_t^+

Lemma E.11. *Assume that $N_{W_t}(\boldsymbol{x}^+)\geq\gamma$ and $-N_{W_t}(\boldsymbol{x}^-)<1$ for $T\leq t<T+b$ where $b\geq2$. Then $N_{W_{T+b}}(\boldsymbol{x}^+)\leq N_{W_T}(\boldsymbol{x}^+)-(b-1)c_\eta+\eta\left|W_0^+(2)\cup W_0^+(4)\right|$.*

Proof. Define $R_t^+=Y_t^+-Z_t^+$ where

$$Y_t^+=\sum_{j\in W_t^+(2)\cup W_t^+(4)}\left[\max\left\{\sigma\left(\boldsymbol{w}^{(i)}\cdot\boldsymbol{x}_1^+\right),...,\sigma\left(\boldsymbol{w}^{(i)}\cdot\boldsymbol{x}_d^+\right)\right\}\right]$$

and

$$Z_t^+=\sum_{j\in U_t^+(2)\cup U_t^+(4)}\left[\max\left\{\sigma\left(\boldsymbol{u}^{(i)}\cdot\boldsymbol{x}_1^+\right),...,\sigma\left(\boldsymbol{u}^{(i)}\cdot\boldsymbol{x}_d^+\right)\right\}\right]$$

Let $l\in\{2,4\}$, $t=T$ and $j\in U_{t+1}^+(l)$. Then, either $j\in U_t^+(2)\cup U_t^+(4)$ or $j\in U_t^+(1)\cup U_t^+(3)$. In the first case, $\boldsymbol{u}_{t+1}^{(j)}=\boldsymbol{u}_t^{(j)}+\eta\boldsymbol{p}_l$. Note that this implies that $U_t^+(2)\cup U_t^+(4)\subseteq U_{t+1}^+(2)\cup U_{t+1}^+(4)$

(since $\boldsymbol{p}_l$ will remain the maximal direction). Therefore,

$$\begin{aligned}
&\sum_{j \in \left(U_{t+1}^+(2) \cup U_{t+1}^+(4)\right) \bigcap \left(U_t^+(2) \cup U_t^+(4)\right)} \left[\max\left\{\sigma\left(\boldsymbol{u}_{t+1}^{(j)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{u}_{t+1}^{(j)} \cdot \boldsymbol{x}_d^+\right)\right\}\right] \\
&- \sum_{j \in U_t^+(2) \cup U_t^+(4)} \left[\max\left\{\sigma\left(\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{u}_{t+1}^{(j)} \cdot \boldsymbol{x}_d^+\right)\right\}\right] \\
&= \eta \left| \left(U_{t+1}^+(2) \cup U_{t+1}^+(4)\right) \bigcap \left(U_t^+(2) \cup U_t^+(4)\right) \right| \\
&= \eta \left| U_t^+(2) \cup U_t^+(4) \right| \qquad (6)
\end{aligned}$$

In the second case, where we have $j \in U_t^+(1) \cup U_t^+(3)$, it holds that $\boldsymbol{u}_{t+1}^{(j)} = \boldsymbol{u}_t^{(j)} + \eta \boldsymbol{p}_l$, $j \in U_t^-(l)$ and $\boldsymbol{u}_{t+1}^{(j)} \cdot \boldsymbol{p}_l > \eta$. Furthermore, by Lemma E.7, $\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{p}_i < \eta$ for $i \in \{1, 3\}$. Note that by Lemma E.7, any $j_1 \in U_t^+(1) \cup U_t^+(3)$ satisfies $j_1 \in U_{t+1}^+(2) \cup U_{t+1}^+(4)$. By all these observations, we have

$$\begin{aligned}
&\sum_{j \in \left(U_{t+1}^+(2) \cup U_{t+1}^+(4)\right) \bigcap \left(U_t^+(1) \cup U_t^+(3)\right)} \left[\max\left\{\sigma\left(\boldsymbol{u}_{t+1}^{(j)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{u}_{t+1}^{(j)} \cdot \boldsymbol{x}_d^+\right)\right\}\right] \\
&- \sum_{j \in U_t^+(1) \cup U_t^+(3)} \left[\max\left\{\sigma\left(\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{u}_{t+1}^{(j)} \cdot \boldsymbol{x}_d^+\right)\right\}\right] \\
&\geq 0 \qquad (7)
\end{aligned}$$

By Eq. 6 and Eq. 7, it follows that, $Z_{t+1}^+ + P_{t+1}^+ \geq Z_{t+1}^+ \geq Z_t^+ + P_t^+ + \eta \left| U_t^+(2) \cup U_t^+(4) \right|$. By induction we have $Z_{t+b}^+ + P_{t+b}^+ \geq Z_t^+ + P_t^+ + \sum_{i=0}^{b-1} \eta \left| U_{t+i}^+(2) \cup U_{t+i}^+(4) \right|$. By Lemma E.7 for any $1 \leq i \leq b-1$ we have $\left| U_{t+i}^+(2) \cup U_{t+i}^+(4) \right| = \{1, ..., k\}$. Therefore, $Z_{t+b}^+ + P_{t+b}^+ \geq Z_t^+ + P_t^+ + (b-1)c_\eta$.

Now, assume that $j \in W_T^+(l)$ for $l \in \{2, 4\}$. Then $\boldsymbol{w}_{T+1}^{(j)} = \boldsymbol{w}_T^{(j)} - \eta \boldsymbol{p}_l$. Thus either

$$\max\left\{\sigma\left(\boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{x}_d^+\right)\right\} - \max\left\{\sigma\left(\boldsymbol{w}_T^{(j)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{w}_T^{(j)} \cdot \boldsymbol{x}_d^+\right)\right\} = -\eta$$

in the case that $j \in W_{T+1}^+(l)$, or

$$\max\left\{\sigma\left(\boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{x}_d^+\right)\right\} \leq \eta$$

if $j \notin W_{T+1}^+(l)$.

Applying these observations b times, we see that $Y_{T+b}^+ - Y_T^+$ is at most $\eta \left| W_{T+b}^+(2) \cup W_{T+b}^+(4) \right| = \eta \left| W_0^+(2) \cup W_0^+(4) \right|$ where the equality follows by Lemma E.5. By Lemma E.10, we have $S_{T+b}^+ = S_T^+$.

Hence, we can conclude that

$$\begin{aligned}
N_{W_{T+b}}(\boldsymbol{x}^+) - N_{W_T}(\boldsymbol{x}^+) &= S_{T+b}^+ + R_{T+b}^+ - P_{T+b}^+ - S_T^- - R_T^+ + P_T^+ \\
&= Y_{T+b}^+ - Z_{T+b}^+ - P_{T+b}^+ - Y_T^+ + Z_T^+ + P_T^+ \\
&\leq -(b-1)c_\eta + \eta \left| W_0^+(2) \cup W_0^+(4) \right|
\end{aligned}$$

□

Lemma E.12. *Assume that $N_{W_t}(\boldsymbol{x}^+) < \gamma$ and $-N_{W_t}(\boldsymbol{x}^-) \geq 1$ for $T \leq t < T + b$ where $b \geq 1$. Then $-N_{W_{T+b}}(\boldsymbol{x}^-) \leq -N_{W_T}(\boldsymbol{x}^-) - b\eta \left| W_0^+(2) \cup W_0^+(4) \right| + c_\eta$.*

Proof. Define

$$Y_t^- = \sum_{j \in W_t^+(2) \cup W_t^+(4)} \left[\max\left\{\sigma\left(\boldsymbol{w}^{(i)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{w}^{(i)} \cdot \boldsymbol{x}_d^+\right)\right\}\right]$$

and

$$Z_t^- = \sum_{j=1}^{k} \left[\max\left\{\sigma\left(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}_d^+\right)\right\}\right]$$

First note that by Lemma E.5 we have $W_{t+1}^+(2) \cup W_{t+1}^+(4) = W_t^+(2) \cup W_t^+(4)$. Next, for any $l \in \{2,4\}$ and $j \in W_t^+(l)$ we have $\boldsymbol{w}_{t+1}^{(j)} = \boldsymbol{w}_t^{(j)} + \eta \boldsymbol{p}_l$. Therefore,

$$Y_{T+b}^- \geq Y_T^- + b\eta \left|W_T^+(2) \cup W_T^+(4)\right| = Y_T^- + b\eta \left|W_0^+(2) \cup W_0^+(4)\right|$$

where the second equality follows by Lemma E.5.

Assume that $j \in U_T^+(l)$ for $l \in \{1,3\}$. Then $\boldsymbol{u}_{T+1}^{(j)} = \boldsymbol{u}_T^{(j)} - \eta \boldsymbol{p}_l$ and

$$\max\left\{\sigma\left(\boldsymbol{u}_{T+1}^{(j)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_{T+1}^{(j)} \cdot \boldsymbol{x}_d^-\right)\right\} - \max\left\{\sigma\left(\boldsymbol{u}_T^{(j)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_T^{(j)} \cdot \boldsymbol{x}_d^-\right)\right\} = 0 \tag{8}$$

To see this, note that by Lemma E.7 and Lemma E.6 it holds that $\boldsymbol{u}_T^{(j)} = \boldsymbol{u}_0^{(j)} + a_T \eta \boldsymbol{p}_l$ where $a_T \in \{-1, 0\}$. Hence, $\boldsymbol{u}_{T+1}^{(j)} = \boldsymbol{u}_0^{(j)} + a_{T+1} \eta \boldsymbol{p}_l$ where $a_{T+1} \in \{-1, 0\}$. Since $\left|\boldsymbol{u}_0^{(j)} \cdot \boldsymbol{p}_2\right| < \frac{\sqrt{2}\eta}{4}$ it follows that $\boldsymbol{u}_{T+1}^{(j)} \cdot \boldsymbol{p}_2 = \boldsymbol{u}_T^{(j)} \cdot \boldsymbol{p}_2 = \boldsymbol{u}_0^{(j)} \cdot \boldsymbol{p}_2$ and thus Eq. 8 holds.

Now assume that $j \in U_T^+(l)$ for $l \in \{2,4\}$. Then

$$\max\left\{\sigma\left(\boldsymbol{u}_{T+1}^{(j)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_{T+1}^{(j)} \cdot \boldsymbol{x}_d^-\right)\right\} - \max\left\{\sigma\left(\boldsymbol{u}_T^{(j)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_T^{(j)} \cdot \boldsymbol{x}_d^-\right)\right\} = -\eta$$

if $l \in \{2,4\}$ and $j \in U_{T+1}^+(l)$, or

$$\max\left\{\sigma\left(\boldsymbol{u}_{T+1}^{(j)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_{T+1}^{(j)} \cdot \boldsymbol{x}_d^-\right)\right\} \leq \eta$$

if $l \in \{2,4\}$ and $j \notin U_{T+1}^+(l)$.

Applying these observations b times, we see that $Z_{T+b}^- - Z_T^-$ is at most $\eta \left|U_{T+b}^+(2) \cup U_{T+b}^+(4)\right|$. Furthermore, for $j \in W_T^+(l)$, $l \in \{1,3\}$, it holds that $\boldsymbol{w}_{T+1}^{(j)} = \boldsymbol{w}_T^{(j)} + \eta \boldsymbol{p}_l$. Therefore

$$\max\left\{\sigma\left(\boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{w}_{T+1}^{(j)} \cdot \boldsymbol{x}_d^-\right)\right\} = \max\left\{\sigma\left(\boldsymbol{w}_T^{(j)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{w}_T^{(j)} \cdot \boldsymbol{x}_d^-\right)\right\}$$

and since $W_{T+1}^+(1) \cup W_{T+1}^+(3) = W_T^+(1) \cup W_T^+(3)$ by Lemma E.5, we get $S_{T+b}^- = S_T^-$. Hence, we can conclude that

$$\begin{aligned}
-N_{W_{T+b}}(\boldsymbol{x}^-) + N_{W_T}(\boldsymbol{x}^-) &= -S_{T+b}^- - Y_{T+b}^- + Z_{T+b}^- + S_T^- + Y_T^- - Z_T^- \\
&\leq -b\eta \left|W_0^+(2) \cup W_0^+(4)\right| + \eta \left|U_{T+b}^+(2) \cup U_{T+b}^+(4)\right| \\
&\leq -b\eta \left|W_0^+(2) \cup W_0^+(4)\right| + c_\eta
\end{aligned}$$

□

Lemma E.13. *For all t, $N_{W_t}(\boldsymbol{x}^+) \leq \gamma + 3c_\eta$, $-N_{W_t}(\boldsymbol{x}^-) \leq 1 + 3c_\eta$ and $S_t^+ \leq \gamma + 1 + 8c_\eta$.*

Proof. The claim holds for $t = 0$. Consider an iteration T. If $N_{W_T}(\boldsymbol{x}^+) < \gamma$ then $N_{W_{T+1}}(\boldsymbol{x}^+) \leq N_{W_T}(\boldsymbol{x}^+) + 2\eta k \leq \gamma + 2c_\eta$. Now assume that $N_{W_t}(\boldsymbol{x}^+) \geq \gamma$ for $T \leq t \leq T + b$ and $N_{W_{T-1}}(\boldsymbol{x}^+) < \gamma$. By Lemma E.11, it holds that $N_{W_{T+b}}(\boldsymbol{x}^+) \leq N_{W_T}(\boldsymbol{x}^+) + \eta k \leq N_{W_T}(\boldsymbol{x}^+) + c_\eta \leq \gamma + 3c_\eta$, where the last inequality follows from the previous observation. Hence, $N_{W_t}(\boldsymbol{x}^+) \leq \gamma + 3c_\eta$ for all t.

The proof of the second claim follows similarly. It holds that $-N_{W_{T+1}}(\boldsymbol{x}^-) < 1 + 2c_\eta$ if $-N_{W_T}(\boldsymbol{x}^-) < 1$. Otherwise if $-N_{W_t}(\boldsymbol{x}^-) \geq 1$ for $T \leq t \leq T + b$ and $-N_{W_{T-1}}(\boldsymbol{x}^-) < 1$ then $-N_{W_{T+b}}(\boldsymbol{x}^-) \leq 1 + 3c_\eta$ by Lemma E.12.

The third claim holds by the following identities and bounds $N_{W_T}(\boldsymbol{x}^+) - N_{W_T}(\boldsymbol{x}^-) = S_T^+ - P_T^+ + P_T^- - S_T^-$, $P_T^- \geq 0$, $\left|P_T^+\right| \leq c_\eta$, $\left|S_T^-\right| \leq c_\eta$ and $N_{W_T}(\boldsymbol{x}^+) - N_{W_T}(\boldsymbol{x}^-) \leq \gamma + 1 + 6c_\eta$ by the previous claims. □

E.0.6 Optimization

We are now ready to prove a global optimality guarantee for gradient descent.

Proposition E.14. *Let $k > 16$ and $\gamma \geq 1$. With probabaility at least $1 - \frac{\sqrt{8k}}{\sqrt{\pi}e^{8k}} - 4e^{-8}$, after $T = \frac{7(\gamma+1+8c_\eta)}{\left(\frac{k}{2}-2\sqrt{k}\right)\eta}$ iterations, gradient descent converges to a global minimum.*

Proof. First note that with probability at least $1 - \frac{\sqrt{8k}}{\sqrt{\pi}e^{8k}} - 4e^{-8}$ the claims of Lemma E.2 and Lemma E.3 hold. Now, if gradient descent has not reached a global minimum at iteration t then either $N_{W_t}(\boldsymbol{x}^+) < \gamma$ or $-N_{W_t}(\boldsymbol{x}^-) < 1$. If $-N_{W_t}(\boldsymbol{x}^+) < \gamma$ then by Lemma E.10 it holds that

$$S_{t+1}^+ \geq S_t^+ + \eta \left|W_0^+(1) \cup W_0^+(3)\right| \geq S_t^+ + \left(\frac{k}{2} - 2\sqrt{k}\right)\eta \tag{9}$$

where the last inequality follows by Lemma E.2.

If $N_{W_t}(\boldsymbol{x}^+) \geq \gamma$ and $-N_{W_t}(\boldsymbol{x}^-) < 1$ we have $S_{t+1}^+ = S_t^+$ by Lemma E.10. However, by Lemma E.11, it follows that after 5 consecutive iterations $t < t' < t + 6$ in which $N_{W_{t'}}(\boldsymbol{x}^+) \geq \gamma$ and $-N_{W_{t'}}(\boldsymbol{x}^-) < 1$, we have $N_{W_{t+6}}(\boldsymbol{x}^+) < \gamma$. To see this, first note that for all t, $N_{W_t}(\boldsymbol{x}^+) \leq \gamma + 3c_\eta$ by Lemma E.13. Then, by Lemma E.11 we have

$$\begin{aligned} N_{W_{t+6}}(\boldsymbol{x}^+) &\leq N_{W_t}(\boldsymbol{x}^+) - 5c_\eta + \eta \left|W_0^+(2) \cup W_0^+(4)\right| \\ &\leq \gamma + 3c_\eta - 5c_\eta + c_\eta \\ &< \gamma \end{aligned}$$

where the second inequality follows by Lemma E.2 and the last inequality by the assumption on k.

Assume by contradiction that GD has not converged to a global minimum after $T = \frac{7(\gamma+1+8c_\eta)}{\left(\frac{k}{2}-2\sqrt{k}\right)\eta}$ iterations. Then, by the above observations, and the fact that $S_0^+ > 0$ with probability 1, we have

$$\begin{aligned} S_T^+ &\geq S_0^+ + \left(\frac{k}{2} - 2\sqrt{k}\right)\eta\frac{T}{7} \\ &> \gamma + 1 + 8c_\eta \end{aligned}$$

However, this contradicts Lemma E.13. □

E.0.7 Generalization on Positive Class

We will first need the following three lemmas.

Lemma E.15. *With probability at least $1 - 4e^{-8}$, it holds that*

$$\left|\left|W_0^+(1)\right| - \frac{k}{4}\right| \leq 2\sqrt{k}$$

and

$$\left|\left|W_0^+(3)\right| - \frac{k}{4}\right| \leq 2\sqrt{k}$$

Proof. The proof is similar to the proof of Lemma E.2. □

Lemma E.16. *Assume that gradient descent converged to a global minimum at iteration T. Then there exists an iteration $T_2 < T$ for which $S_t^+ \geq \gamma + 1 - 3c_\eta$ for all $t \geq T_2$ and for all $t < T_2$, $-N_{W_t}(\boldsymbol{x}^-) < 1$.*

Proof. Assume that for all $0 \leq t \leq T_1$ it holds that $N_{W_t}(\boldsymbol{x}^+) < \gamma$ and $-N_{W_t}(\boldsymbol{x}^-) < 1$. By continuing the calculation of Lemma E.4 we have the following:

1. For $i \in \{1,3\}$, $l \in \{2,4\}$, $j \in W_0^+(i) \cap W_0^-(l)$, it holds that $\boldsymbol{w}_{T_1}^{(j)} = \boldsymbol{w}_0^{(j)} + T_1\eta\boldsymbol{p}_i - \frac{1}{2}(1-(-1)^{T_1})\eta\boldsymbol{p}_l$.
2. For $i \in \{2,4\}$ and $j \in W_0^+(i)$, it holds that $\boldsymbol{w}_{T_1}^{(j)} = \boldsymbol{w}_0^{(j)}$.
3. For $i \in \{1,3\}$, $l \in \{2,4\}$, $j \in U_0^+(i) \cap U_0^-(l)$, it holds that $\boldsymbol{u}_{T_1}^{(j)} = \boldsymbol{u}_0^{(j)} - \eta\boldsymbol{p}_i + \eta\boldsymbol{p}_l$.
4. For $i \in \{2,4\}$ and $j \in U_0^+(i)$, it holds that $\boldsymbol{u}_{T_1}^{(j)} = \boldsymbol{u}_0^{(j)}$.

Therefore, there exists an iteration T_1 such that $N_{W_{T_1}}(\boldsymbol{x}^+) \geq \gamma$ and $-N_{W_{T_1}}(\boldsymbol{x}^-) < 1$ and for all $t < T_1$, $N_{W_t}(\boldsymbol{x}^+) < \gamma$ and $-N_{W_t}(\boldsymbol{x}^-) < 1$. Let $T_2 \leq T$ be the first iteration such that $-N_{W_{T_2}}(\boldsymbol{x}^-) \geq 1$. We claim that for all $T_1 \leq t \leq T_2$ we have $N_{W_{T_1}}(\boldsymbol{x}^+) \geq \gamma - 2c_\eta$. It suffices to show that for all $T_1 \leq t < T_2$ the following holds:

1. If $N_{W_t}(\boldsymbol{x}^+) \geq \gamma$ then $N_{W_{t+1}}(\boldsymbol{x}^+) \geq \gamma - 2c_\eta$.
2. If $N_{W_t}(\boldsymbol{x}^+) < \gamma$ then $N_{W_{t+1}}(\boldsymbol{x}^+) \geq N_{W_t}(\boldsymbol{x}^+)$.

The first claim follows since at any iteration $N_{W_t}(\boldsymbol{x}^+)$ can decrease by at most $2\eta k = 2c_\eta$. For the second claim, let $t' < t$ be the latest iteration such that $N_{W_{t'}}(\boldsymbol{x}^+) \geq \gamma$. Then at iteration t' it holds that $-N_{W_{t'}}(\boldsymbol{x}^-) < 1$ and $N_{W_{t'}}(\boldsymbol{x}^+) \geq \gamma$. Therefore, for all $i \in \{1,3\}$, $l \in \{2,4\}$ and $j \in U_0^+(i) \cap U_0^+(l)$ it holds that $\boldsymbol{u}_{t'+1}^{(j)} = \boldsymbol{u}_{t'}^{(j)} + \eta\boldsymbol{p}_l$. Hence, by Lemma E.6 and Lemma E.7 it holds that $U_{t'+1}^+(1) \cup U_{t'+1}^+(3) = \emptyset$. Therefore, by the gradient update in Eq. 3, for all $1 \leq j \leq k$, and all $t' < t'' \leq t$ we have $\boldsymbol{u}_{t''+1}^{(j)} = \boldsymbol{u}_{t''}^{(j)}$, which implies that $N_{W_{t''+1}}(\boldsymbol{x}^+) \geq N_{W_{t''}}(\boldsymbol{x}^+)$. For $t'' = t$ we get $N_{W_{t+1}}(\boldsymbol{x}^+) \geq N_{W_t}(\boldsymbol{x}^+)$.

The above argument shows that $N_{W_{T_2}}(\boldsymbol{x}^+) \geq \gamma - 2c_\eta$ and $-N_{W_{T_2}}(\boldsymbol{x}^-) \geq 1$. Since $N_{W_{T_2}}(\boldsymbol{x}^+) - N_{W_{T_2}}(\boldsymbol{x}^-) = S_{T_2}^+ - P_{T_2}^+ + P_{T_2}^- - S_{T_2}^-$, $P_{T_2}^-, S_{T_2}^- \geq 0$ and $\left|P_{T_2}^-\right| \leq c_\eta$ it follows that $S_{T_2}^+ \geq \gamma + 1 - 3c_\eta$. Finally, by Lemma E.10 we have $S_t^+ \geq \gamma + 1 - 3c_\eta$ for all $t \geq T_2$. □

Lemma E.17. *Let*

$$X_t^+ = \sum_{j \in W_t^+(2) \cup W_t^+(4)} \left[\max\left\{\sigma\left(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}_d^+\right)\right\}\right]$$

and

$$Y_t^+ = \sum_{j \in U_t^+(2) \cup U_t^+(4)} \left[\max\left\{\sigma\left(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}_d^+\right)\right\}\right]$$

Assume that $k \geq 64$ and gradient descent converged to a global minimum at iteration T. Then, $X_T^+ \leq 34c_\eta$ and $Y_T^+ \leq 1 + 38c_\eta$.

Proof. Notice that by the gradient update in Eq. 2 and Lemma E.3, X_t^+ can be strictly larger than $\max\left\{X_{t-1}^+, \eta\left|W_t^+(2) \cup W_t^+(4)\right|\right\}$ only if $N_{W_{t-1}}(\boldsymbol{x}^+) < \gamma$ and $-N_{W_{t-1}}(\boldsymbol{x}^-) \geq 1$. Furthermore, in this case $X_t^+ - X_{t-1}^+ = \eta\left|W_t^+(2) \cup W_t^+(4)\right|$. By Lemma E.10, S_t^+ increases in this case by $\eta\left|W_t^+(1) \cup W_t^+(3)\right|$. We know by Lemma E.16 that there exists $T_2 < T$ such that $S_{T_2}^+ \geq \gamma + 1 - 3c_\eta$ and that $N_{W_t}(\boldsymbol{x}^+) < \gamma$ and $-N_{W_t}(\boldsymbol{x}^-) \geq 1$ only for $t > T_2$. Since $S_t^+ \leq \gamma + 1 + 8c_\eta$ for all t by Lemma E.13, there can only be at most $\frac{11c_\eta}{\eta\left|W_T^+(1) \cup W_T^+(3)\right|}$ iterations in which $N_{W_t}(\boldsymbol{x}^+) < \gamma$ and $-N_{W_t}(\boldsymbol{x}^-) \geq 1$. It follows that

$$\begin{aligned}
X_t^+ &\leq \eta\left|W_T^+(2) \cup W_T^+(4)\right| + \frac{11c_\eta\eta\left|W_T^+(2) \cup W_T^+(4)\right|}{\eta\left|W_T^+(1) \cup W_T^+(3)\right|} \\
&\leq c_\eta + 11c_\eta\frac{\left(\frac{k}{2} + 2\sqrt{k}\right)}{\left(\frac{k}{2} - 2\sqrt{k}\right)} \\
&\leq 34c_\eta
\end{aligned}$$

where the second inequality follows by Lemma E.2 and the third inequality by the assumption on k.

At convergence we have $N_{W_T}(\boldsymbol{x}^-) = S_T^- + X_T^+ - Y_T^+ - P_T^- \geq -1 - 3c_\eta$ by Lemma E.13 (recall that $R_t^- = R_t^+ = X_t^+ - Y_t^+$). Furthermore, $P_T^- \geq 0$ and by Lemma E.8 we have $S_T^- \leq c_\eta$. Therefore, we get $Y_T^+ \leq 1 + 38c_\eta$. □

We are now ready to prove the main result of this section.

Proposition E.18. *Define* $\beta(\gamma) = \frac{\gamma - 40\frac{1}{4}c_\eta}{39c_\eta + 1}$. *Assume that* $\gamma \geq 2$ *and* $k \geq 64\left(\frac{\beta(\gamma)+1}{\beta(\gamma)-1}\right)^2$. *Then with probability at least* $1 - \frac{\sqrt{8k}}{\sqrt{\pi}e^{8k}} - 8e^{-8}$, *gradient descent converges to a global minimum which classifies all positive points correctly.*

Proof. With probability at least $1 - \frac{\sqrt{128k}}{\sqrt{\pi}e^{\frac{k}{2}}} - 8e^{-8}$ Proposition E.14, and Lemma E.15 hold. It suffices to show generalization on positive points. Assume that gradient descent converged to a global minimum at iteration T. Let $(\boldsymbol{z}, 1)$ be a positive point. Then there exists $\boldsymbol{z}_i \in \{(1,1), (-1,-1)\}$. Assume without loss of generality that $\boldsymbol{z}_i = (-1,-1) = \boldsymbol{p}_3$. Define

$$X_t^+(i) = \sum_{j \in W_T^+(i)} \left[\max\left\{\sigma\left(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}_d^+\right)\right\}\right]$$

$$Y_t^+(i) = \sum_{j \in U_T^+(i)} \left[\max\left\{\sigma\left(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}_1^+\right), ..., \sigma\left(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}_d^+\right)\right\}\right]$$

for $i \in [4]$.

Notice that

$$\begin{aligned} N_{W_T}(\boldsymbol{x}^+) &= X_T^+(1) + X_T^+(3) - P_T^+ + R_T^+ \\ &= X_T^+(1) + X_T^+(3) - P_T^+ + R_T^- \\ &= X_T^+(1) + X_T^+(3) - P_T^+ + N_{W_T}(\boldsymbol{x}^-) - S_T^- + P_T^- \end{aligned}$$

Since $N_{W_T}(\boldsymbol{x}^+) \geq \gamma$, $-N_{W_T}(\boldsymbol{x}^-) \geq 1$, $\left|P_T^-\right| \leq c_\eta$ by Lemma E.8 and $P_T^+, S_T^- \geq 0$, we obtain

$$X_T^+(1) + X_T^+(3) \geq \gamma + 1 - c_\eta \tag{10}$$

Furthermore, by Lemma E.9 we have

$$\frac{X_T^+(1) - X_0^+(1)}{\left|W_T^+(1)\right|} = \frac{X_T^+(3) - X_0^+(3)}{\left|W_T^+(3)\right|} \tag{11}$$

and by Lemma E.15,

$$\frac{\frac{k}{4} - 2\sqrt{k}}{\frac{k}{4} + 2\sqrt{k}} \leq \frac{\left|W_T^+(1)\right|}{\left|W_T^+(3)\right|} \leq \frac{\frac{k}{4} + 2\sqrt{k}}{\frac{k}{4} - 2\sqrt{k}} \tag{12}$$

Let $\alpha(k) = \frac{\frac{k}{4} + 2\sqrt{k}}{\frac{k}{4} - 2\sqrt{k}}$. By Lemma E.3 we have $\left|X_0^+(1)\right| \leq \frac{\eta k}{4} \leq \frac{c_\eta}{4}$. Combining this fact with Eq. 11 and Eq. 12 we get

$$X_T^+(1) \leq \alpha(k) X_T^+(3) + X_0^+(1) \leq \alpha(k) X_T^+(3) + \frac{c_\eta}{4}$$

which implies together with Eq. 10 that $X_T^+(3) \geq \frac{\gamma + 1 - \frac{5c_\eta}{4}}{1 + \alpha(k)}$. Therefore,

$$\begin{aligned} N_{W_T}(\boldsymbol{z}) &\geq X_T^+(3) - P_T^+ - Y_T^+(2) - Y_T^+(4) \\ &\geq \frac{\gamma + 1 - \frac{5c_\eta}{4}}{1 + \alpha(k)} - c_\eta - 1 - 3(8c_\eta) - 14c_\eta \\ &= \frac{\gamma + 1 - \frac{5c_\eta}{4}}{1 + \alpha(k)} - 39c_\eta - 1 > 0 \end{aligned} \tag{13}$$

where the first inequality is true because

$$\sum_{j=1}^{k}\left[\max\left\{\sigma\left(\boldsymbol{u}^{(j)}\cdot\boldsymbol{z}_1\right),...,\sigma\left(\boldsymbol{u}^{(j)}\cdot\boldsymbol{z}_d\right)\right\}\right]\leq\sum_{j=1}^{k}\left[\max\left\{\sigma\left(\boldsymbol{u}^{(j)}\cdot\boldsymbol{x}_1^{+}\right),...,\sigma\left(\boldsymbol{u}^{(j)}\cdot\boldsymbol{x}_d^{+}\right)\right\}\right] \tag{14}$$

$$=P_T^{+}+Y_T^{+}(2)+Y_T^{+}(4) \tag{15}$$

The second inequality in Eq. 13 follows since $P_T^{+}\leq c_\eta$ and by appyling Lemma E.17. Finally, the last inequality in Eq. 13 follows by the assumption on k. [3] Hence, $\boldsymbol{z}$ is classified correctly. □

E.0.8 Generalization on Negative Class

We will need the following lemmas.

Lemma E.19. *With probability at least* $1-8e^{-8}$*, it holds that*

$$\left|\left|U_0^{+}(2)\right|-\frac{k}{4}\right|\leq 2\sqrt{k}$$

$$\left|\left|U_0^{+}(4)\right|-\frac{k}{4}\right|\leq 2\sqrt{k}$$

$$\left|\left|\left(U_0^{+}(1)\cup U_0^{+}(3)\right)\cap U_0^{-}(2)\right|-\frac{k}{4}\right|\leq 2\sqrt{k}$$

$$\left|\left|\left(U_0^{+}(1)\cup U_0^{+}(3)\right)\cap U_0^{-}(4)\right|-\frac{k}{4}\right|\leq 2\sqrt{k}$$

Proof. The proof is similar to the proof of Lemma E.2 and follows from the fact that

$$\begin{aligned}\mathbb{P}\left[j\in U_0^{+}(2)\right]&=\mathbb{P}\left[j\in U_0^{+}(4)\right]\\&=\mathbb{P}\left[j\in\left(U_0^{+}(1)\cup U_0^{+}(3)\right)\cap U_0^{-}(2)\right]\\&=\mathbb{P}\left[j\in\left(U_0^{+}(1)\cup U_0^{+}(3)\right)\cap U_0^{-}(4)\right]\\&=\frac{1}{4}\end{aligned}$$

□

Lemma E.20. *Let*

$$X_t^{-}=\sum_{j\in U_0^{+}(2)}\left[\max\left\{\sigma\left(\boldsymbol{u}_t^{(j)}\cdot\boldsymbol{x}_1^{-}\right),...,\sigma\left(\boldsymbol{u}_t^{(j)}\cdot\boldsymbol{x}_d^{-}\right)\right\}\right]$$

and

$$Y_t^{-}=\sum_{j\in U_0^{+}(4)}\left[\max\left\{\sigma\left(\boldsymbol{u}_t^{(j)}\cdot\boldsymbol{x}_1^{-}\right),...,\sigma\left(\boldsymbol{u}_t^{(j)}\cdot\boldsymbol{x}_d^{-}\right)\right\}\right]$$

Then for all t*, there exists* $X,Y\geq 0$ *such that* $|X|\leq\eta\left|U_0^{+}(2)\right|$, $|Y|\leq\eta\left|U_0^{+}(4)\right|$ *and* $\frac{X_t^{-}-X}{\left|U_0^{+}(2)\right|}=\frac{Y_t^{-}-Y}{\left|U_0^{+}(4)\right|}$.

Proof. First, we will prove that for all t there exists $a_t\in\mathbb{Z}$ such that for $j_1\in U_0^{-}(2)$ and $j_2\in U_0^{-}(4)$ it holds that $\boldsymbol{u}_t^{(j_1)}=\boldsymbol{u}_0^{(j_1)}+a_t\eta\boldsymbol{p}_2$ and $\boldsymbol{u}_t^{(j_2)}=\boldsymbol{u}_0^{(j_2)}-a_t\eta\boldsymbol{p}_2$. [4] We will prove this by induction on t.

[3] The inequality $\frac{\gamma+1-\frac{5c_\eta}{4}}{1+\alpha(k)}-39c_\eta-1>0$ is equivalent to $\alpha(k)<\beta(\gamma)$ which is equivalent to $k>64\left(\frac{\beta(\gamma)+1}{\beta(\gamma)-1}\right)^2$.

[4] Recall that by Lemma E.6 we know that $U_0^{+}(2)\cup U_0^{+}(4)\subseteq U_t^{+}(2)\cup U_t^{+}(4)$.

For $t = 0$ this clearly holds. Assume it holds for an iteration t. Let $j_1 \in U_0^-(2)$ and $j_2 \in U_0^-(4)$. By the induction hypothesis, there exists $a_T \in \mathbb{Z}$ such that $\boldsymbol{u}_t^{(j_1)} = \boldsymbol{u}_0^{(j_1)} + a_t\eta\boldsymbol{p}_2$ and $\boldsymbol{u}_t^{(j_2)} = \boldsymbol{u}_0^{(j_2)} - a_t\eta\boldsymbol{p}_2$. Since for all $1 \leq j \leq k$ it holds that $\left|\boldsymbol{u}_0^{(j)} \cdot \boldsymbol{p}_2\right| < \frac{\sqrt{2}\eta}{4}$, it follows that either $U_0^-(2) \subseteq U_t^-(2)$ and $U_0^-(4) \subseteq U_t^-(4)$ or $U_0^-(2) \subseteq U_t^-(4)$ and $U_0^-(4) \subseteq U_t^-(2)$. In either case, by Eq. 3, we have the following update at iteration $t+1$:

$$\boldsymbol{u}_{t+1}^{(j_1)} = \boldsymbol{u}_t^{(j_1)} + a\eta\boldsymbol{p}_2$$

and

$$\boldsymbol{u}_{t+1}^{(j_2)} = \boldsymbol{u}_t^{(j_2)} - a\eta\boldsymbol{p}_2$$

where $a \in \{-1, 0, 1\}$. Hence, $\boldsymbol{u}_{t+1}^{(j_1)} = \boldsymbol{u}_0^{(j_1)} + (a_t + a)\eta\boldsymbol{p}_2$ and $\boldsymbol{u}_t^{(j_2)} = \boldsymbol{u}_0^{(j_2)} - (a_t + a)\eta\boldsymbol{p}_2$. This concludes the proof by induction.

Now, consider an iteration t, $j_1 \in U_0^+(2)$, $j_2 \in U_0^+(4)$ and the integer a_t defined above. If $a_t \geq 0$ then

$$\max\left\{\sigma\left(\boldsymbol{u}_t^{(j_1)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_t^{(j_1)} \cdot \boldsymbol{x}_d^-\right)\right\} - \max\left\{\sigma\left(\boldsymbol{u}_0^{(j_1)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_0^{(j_1)} \cdot \boldsymbol{x}_d^-\right)\right\} = \eta a_t$$

and

$$\max\left\{\sigma\left(\boldsymbol{u}_t^{(j_2)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_t^{(j_2)} \cdot \boldsymbol{x}_d^-\right)\right\} - \max\left\{\sigma\left(\boldsymbol{u}_0^{(j_2)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_0^{(j_2)} \cdot \boldsymbol{x}_d^-\right)\right\} = \eta a_t$$

Define $X = X_0^-$ and $Y = Y_0^-$ then $|X| \leq \eta\left|U_0^-(2)\right|$, $|Y| \leq \eta\left|U_0^-(4)\right|$ and

$$\frac{X_t^- - X}{\left|U_0^-(2)\right|} = \frac{\left|U_0^-(2)\right|\eta a_t}{\left|U_0^-(2)\right|} = \eta a_t = \frac{\left|U_0^-(4)\right|\eta a_t}{\left|U_0^-(4)\right|} = \frac{Y_t^- - Y}{\left|U_0^-(4)\right|}$$

which proves the claim in the case that $a_t \geq 0$.

If $a_t < 0$ it holds that

$$\max\left\{\sigma\left(\boldsymbol{u}_t^{(j_1)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_t^{(j_1)} \cdot \boldsymbol{x}_d^-\right)\right\} - \max\left\{\sigma\left(\left(\boldsymbol{u}_0^{(j_1)} - \boldsymbol{p}_2\right) \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\left(\boldsymbol{u}_0^{(j_1)} - \boldsymbol{p}_2\right) \cdot \boldsymbol{x}_d^-\right)\right\} = \eta(-a_t - 1)$$

and

$$\max\left\{\sigma\left(\boldsymbol{u}_t^{(j_2)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_t^{(j_2)} \cdot \boldsymbol{x}_d^-\right)\right\} - \max\left\{\sigma\left(\left(\boldsymbol{u}_0^{(j_2)} + \boldsymbol{p}_2\right) \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\left(\boldsymbol{u}_0^{(j_2)} + \boldsymbol{p}_2\right) \cdot \boldsymbol{x}_d^-\right)\right\} = \eta(-a_t - 1)$$

Define

$$X = \sum_{j \in U_0^+(2)} \left[\max\left\{\sigma\left(\left(\boldsymbol{u}_0^{(j)} - \boldsymbol{p}_2\right) \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\left(\boldsymbol{u}_0^{(j)} - \boldsymbol{p}_2\right) \cdot \boldsymbol{x}_d^-\right)\right\}\right]$$

and

$$Y = \sum_{j \in U_0^+(4)} \left[\max\left\{\sigma\left(\left(\boldsymbol{u}_0^{(j)} + \boldsymbol{p}_2\right) \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\left(\boldsymbol{u}_0^{(j)} + \boldsymbol{p}_2\right) \cdot \boldsymbol{x}_d^-\right)\right\}\right]$$

Since for all $1 \leq j \leq k$ it holds that $\left|\boldsymbol{u}_0^{(j)} \cdot \boldsymbol{p}_2\right| < \frac{\sqrt{2}\eta}{4}$, we have $|X| \leq \eta\left|U_0^-(2)\right|$, $|Y| \leq \eta\left|U_0^-(4)\right|$. Furthermore,

$$\frac{X_t^- - X}{\left|U_0^-(2)\right|} = \frac{\left|U_0^-(2)\right|\eta(-a_t - 1)}{\left|U_0^-(2)\right|} = \eta(-a_t - 1) = \frac{\left|U_0^-(4)\right|\eta(-a_t - 1)}{\left|U_0^-(4)\right|} = \frac{Y_t^- - Y}{\left|U_0^-(4)\right|}$$

which concludes the proof. □

Lemma E.21. *Let*

$$X_t^- = \sum_{j \in \left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(2)} \left[\max\left\{\sigma\left(\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{x}_d^-\right)\right\}\right]$$

and

$$Y_t^- = \sum_{j \in \left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(4)} \left[\max\left\{\sigma\left(\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{x}_d^-\right)\right\}\right]$$

Then for all t, $\frac{X_t^- - X_0^-}{\left|\left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(2)\right|} = \frac{Y_t^- - Y_t^-}{\left|\left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(4)\right|}$.

Proof. We will first prove that for all t there exists an integer $a_t \geq 0$ such that for $j_1 \in \left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(2)$ and $j_2 \in \left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(4)$ it holds that $\boldsymbol{u}_t^{(j_1)} \cdot \boldsymbol{p}_2 = \boldsymbol{u}_0^{(j_1)} \cdot \boldsymbol{p}_2 + \eta a_t$ and $\boldsymbol{u}_t^{(j_2)} \cdot \boldsymbol{p}_4 = \boldsymbol{u}_0^{(j_2)} \cdot \boldsymbol{p}_4 + \eta a_t$. We will prove this by induction on t.

For $t = 0$ this clearly holds. Assume it holds for an iteration t. Let $j_1 \in \left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(2)$ and $j_2 \in \left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(4)$. By the induction hypothesis, there exists an integer $a_t \geq 0$ such that $\boldsymbol{u}_t^{(j_1)} \cdot \boldsymbol{p}_2 = \boldsymbol{u}_0^{(j_1)} \cdot \boldsymbol{p}_2 + \eta a_t$ and $\boldsymbol{u}_t^{(j_2)} \cdot \boldsymbol{p}_4 = \boldsymbol{u}_0^{(j_2)} \cdot \boldsymbol{p}_4 + \eta a_t$. Since for all $1 \leq j \leq k$ it holds that $\left|\boldsymbol{u}_0^{(j)} \cdot \boldsymbol{p}_1\right| < \frac{\sqrt{2}\eta}{4}$, it follows that if $a_t \geq 1$ we have the following update at iteration $T+1$:

$$\boldsymbol{u}_{t+1}^{(j_1)} = \boldsymbol{u}_t^{(j_1)} + a\eta \boldsymbol{p}_2$$

and

$$\boldsymbol{u}_{t+1}^{(j_2)} = \boldsymbol{u}_t^{(j_2)} + a\eta \boldsymbol{p}_4$$

where $a \in \{-1, 0, 1\}$. Hence, $\boldsymbol{u}_{t+1}^{(j_1)} \cdot \boldsymbol{p}_2 = \boldsymbol{u}_0^{(j_1)} \cdot \boldsymbol{p}_2 + \eta(a_t + a)$ and $\boldsymbol{u}_{t+1}^{(j_2)} \cdot \boldsymbol{p}_4 = \boldsymbol{u}_0^{(j_2)} \cdot \boldsymbol{p}_4 + \eta(a_t + a)$.

Otherwise, if $a_t = 0$ then

$$\boldsymbol{u}_{t+1}^{(j_1)} = \boldsymbol{u}_t^{(j_1)} + a\eta \boldsymbol{p}_2 + b_1 \boldsymbol{p}_1$$

and

$$\boldsymbol{u}_{t+1}^{(j_2)} = \boldsymbol{u}_t^{(j_2)} + a\eta \boldsymbol{p}_4 + b_2 \boldsymbol{p}_1$$

such that $a \in \{0, 1\}$ and $b_1, b_2 \in \{-1, 0, 1\}$. Hence, $\boldsymbol{u}_{t+1}^{(j_1)} \cdot \boldsymbol{p}_2 = \boldsymbol{u}_0^{(j_1)} \cdot \boldsymbol{p}_2 + \eta(a_t + a)$ and $\boldsymbol{u}_{t+1}^{(j_2)} \cdot \boldsymbol{p}_4 = \boldsymbol{u}_0^{(j_2)} \cdot \boldsymbol{p}_4 + \eta(a_t + a)$. This concludes the proof by induction.

Now, consider an iteration t, $j_1 \in \left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(2)$ and $j_2 \in \left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(4)$ and the integer a_t defined above. We have,

$$\max\left\{\sigma\left(\boldsymbol{u}_t^{(j_1)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_t^{(j_1)} \cdot \boldsymbol{x}_d^-\right)\right\} - \max\left\{\sigma\left(\boldsymbol{u}_0^{(j_1)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_0^{(j_1)} \cdot \boldsymbol{x}_d^-\right)\right\} = \eta a_t$$

and

$$\max\left\{\sigma\left(\boldsymbol{u}_t^{(j_2)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_t^{(j_2)} \cdot \boldsymbol{x}_d^-\right)\right\} - \max\left\{\sigma\left(\boldsymbol{u}_0^{(j_2)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_0^{(j_2)} \cdot \boldsymbol{x}_d^-\right)\right\} = \eta a_t$$

It follows that

$$\begin{aligned}
\frac{X_t^- - X_0^-}{\left|\left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(2)\right|} &= \frac{\left|\left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(2)\right| \eta a_t}{\left|\left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(2)\right|} \\
&= \eta a_t \\
&= \frac{\left|\left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(4)\right| \eta a_t}{\left|\left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(4)\right|} \\
&= \frac{Y_t^- - Y_0^-}{\left|\left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(4)\right|}
\end{aligned}$$

which concludes the proof. □

We are now ready to prove the main result of this section.

Proposition E.22. *Define $\beta = \frac{1-36\frac{1}{4}c_\eta}{35c_\eta}$. Assume that $k > 64\left(\frac{\beta+1}{\beta-1}\right)^2$. Then with probability at least $1 - \frac{\sqrt{8k}}{\sqrt{\pi}e^{8k}} - 8e^{-8}$, gradient descent converges to a global minimum which classifies all negative points correctly.*

Proof. With probability at least $1-\frac{\sqrt{8k}}{\sqrt{\pi}e^{8k}}-16e^{-8}$ Proposition E.14 and Lemma E.19 hold. It suffices to show generalization on negative points. Assume that gradient descent converged to a global minimum at iteration T. Let $(\boldsymbol{z}, -1)$ be a negative point. Assume without loss of generality that $\boldsymbol{z}_i = \boldsymbol{p}_2$ for all $1 \leq i \leq d$. Define the following sums for $l \in \{2, 4\}$,

$$X_t^- = \sum_{j \in W_t^+(2) \cup W_t^+(4)} \left[\max\left\{\sigma\left(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}_d^-\right)\right\}\right]$$

$$Y_t^-(l) = \sum_{j \in U_0^+(l)} \left[\max\left\{\sigma\left(\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}_t^{(j)} \cdot \boldsymbol{x}_d^-\right)\right\}\right]$$

$$Z_t^-(l) = \sum_{j \in \left(U_0^+(1) \cup U_0^+(3)\right) \cap U_0^-(l)} \left[\max\left\{\sigma\left(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}_d^-\right)\right\}\right]$$

First, we notice that

$$N_{W_T}(\boldsymbol{x}^-) = S_T^- + X_T^- - Y_T^-(2) - Y_T^-(4) - Z_T^-(2) - Z_T^-(4)$$

$$X_T^-, S_T^- \geq 0$$

and

$$N_{W_T}(\boldsymbol{x}^-) \leq -1$$

imply that

$$Y_T^-(2) + Y_T^-(4) + Z_T^-(2) + Z_T^-(4) \geq 1 \tag{16}$$

We note that by the analysis in Lemma E.19, it holds that for any t, $j_1 \in U_0^+(2)$ and $j_2 \in U_0^+(4)$, either $j_1 \in U_t^+(2)$ and $j_2 \in U_t^+(4)$, or $j_1 \in U_t^+(4)$ and $j_2 \in U_t^+(2)$. We assume without loss of generality that $j_1 \in U_T^+(2)$ and $j_2 \in U_T^+(4)$. It follows that in this case $N_{W_T}(\boldsymbol{z}) \leq S_T^- + X_T^- - Z_T^-(2) - Y_T^-(2)$. [5]Otherwise we would replace $Y_T^-(2)$ with $Y_T^-(4)$ and vice versa and continue with the same proof.

Let $\alpha(k) = \frac{\frac{k}{4}+2\sqrt{k}}{\frac{k}{4}-2\sqrt{k}}$. By Lemma E.21 and Lemma E.19

$$Z_T^-(4) \leq \alpha(k) Z_T^-(2) + Z_0^-(2) \leq \alpha(k) Z_T^-(2) + \frac{c_\eta}{4}$$

and by Lemma E.20 and Lemma E.19 there exists $Y \leq c_\eta$ such that:

$$Y_T^-(4) \leq \alpha(k) Y_T^-(2) + Y \leq \alpha(k) Y_T^-(2) + c_\eta$$

Plugging these inequalities in Eq. 16 we get:

$$\alpha(k) Z_T^-(2) + \frac{c_\eta}{4} + \alpha(k) Y_T^-(2) + c_\eta + Y_T^-(2) + Z_T^-(2) \geq 1$$

which implies that

$$Y_T^-(2) + Z_T^-(2) \geq \frac{1 - \frac{5c_\eta}{4}}{\alpha(k) + 1}$$

[5]The fact that we can omit the term $-Z_T^-(4)$ from the latter inequality follows from Lemma E.7.

By Lemma E.17 we have $X_T^- \leq 34c_\eta$. Hence, by using the inequality $S_T^- \leq c_\eta$ we conclude that

$$N_{W_T}(\boldsymbol{z}) \leq S_T^- + X_T^- - Z_T^-(2) - Y_T^-(2) \leq 35c_\eta - \frac{1-\frac{5c_\eta}{4}}{\alpha(k)+1} < 0$$

where the last inequality holds for $k > 64\left(\frac{\beta+1}{\beta-1}\right)^2$. [6] Therefore, $\boldsymbol{z}$ is classified correctly. □

E.0.9 Finishing the Proof

First, for $k \geq 120$, with probability at least $1 - \frac{\sqrt{8k}}{\sqrt{\pi}e^{8k}} - 16e^{-8}$, Proposition E.14, Lemma E.15 and Lemma E.19 hold. Also, for the bound on T, note that in this case $\frac{28(\gamma+1+8c_\eta)}{c_\eta} \geq \frac{7(\gamma+1+8c_\eta)}{\left(\frac{k}{2}-2\sqrt{k}\right)\eta}$. Define $\beta_1 = \frac{\gamma-40\frac{1}{4}c_\eta}{39c_\eta+1}$ and $\beta_2 = \frac{1-36\frac{1}{4}c_\eta}{35c_\eta}$ and let $\beta = \max\{\beta_1, \beta_2\}$. For $\gamma \geq 8$ and $c_\eta \leq \frac{1}{410}$ it holds that $64\left(\frac{\beta+1}{\beta-1}\right)^2 < 120$. By Proposition E.18 and Proposition E.22, it follows that for $k \geq 120$ gradient descent converges to a global minimum which classifies all points correctly.

We will now prove the clustering effect at a global minimum. By Lemma E.16 it holds that $S_T^+ \geq \gamma + 1 - 3c_\eta \geq \gamma - 1$. Therefore, by Lemma E.5 it follows that

$$2\eta(a^+(T)+1)\left|W_0^+(1) \cup W_0^+(3)\right| \geq S_T^+ \geq \gamma - 1$$

and thus $a^+(T) \geq \frac{\gamma-1}{2c_\eta} - 1$. Therefore, for any $j \in W_0^+(i)$ such that $i \in \{1,3\}$, the cosine of the angle between $\boldsymbol{w}_T^{(j)}$ and $\boldsymbol{p}_i$ is at least

$$\frac{(\boldsymbol{w}_0^{(j)} + a^+(T)\eta\boldsymbol{p}_1 + \alpha_i^t\boldsymbol{p}_2)\cdot\boldsymbol{p}_1}{\sqrt{2}(\|\boldsymbol{w}_0^{(j)}\| + \sqrt{2}a^+(T)\eta + \sqrt{2}\eta)} \geq \frac{2a^+(T)}{2a_1(T)+3} \geq \frac{\gamma-1-2c_\eta}{\gamma-1+c_\eta}$$

where we used the triangle inequality and Lemma E.5. The claim follows for $j \in W_0^+(1) \cup W_0^+(3)$.

F Proof of Theorem 6.4

Theorem F.1. ***(Theorem 6.4 restated)*** *Assume that gradient descent runs with parameaters $\eta = \frac{c_\eta}{k}$ where $c_\eta \leq \frac{1}{41}$, $\sigma_g \leq \frac{c_\eta}{16k^{\frac{3}{2}}}$ and $\gamma \geq 1$. Then, with probability at least $(1-c)\frac{33}{48}$, gradient descent converges to a global minimum that does not recover f^*. Furthermore, there exists $1 \leq i \leq 4$ such that the global minimum misclassifies all points $\boldsymbol{x}$ such that $P_{\boldsymbol{x}} = A_i$.*

We refer to Eq. 9 in the proof of Proposition E.14. To show convergence and provide convergence rates of gradient descent, the proof uses Lemma E.2. However, to only show convergence, it suffices to bound the probability that $W_0^+(1) \cup W_0^+(3) \neq \emptyset$ and that the initialization satisfies Lemma E.3. Given that Lemma E.3 holds (with probability at least $1 - \sqrt{\frac{8}{\pi}}e^{-32}$), then $W_0^+(1) \cup W_0^+(3) \neq \emptyset$ holds with probability $\frac{3}{4}$.

By the argument above, with probability at least $\left(1 - \sqrt{\frac{8}{\pi}}e^{-32}\right)\frac{3}{4}$, Lemma E.3 holds with $k=2$ and $W_0^+(1) \cup W_0^+(3) \neq \emptyset$ which implies that gradient descent converges to a global minimum. For the rest of the proof we will condition on the corresponding event. Let T be the iteration in which gradient descent converges to a global minimum. Note that T is a random variable. Denote the network at iteration T by N. For all $\boldsymbol{z} \in \mathbb{R}^{2d}$ denote

$$N(\boldsymbol{z}) = \sum_{j=1}^{2}\left[\max\left\{\sigma\left(\boldsymbol{w}^{(j)}\cdot\boldsymbol{z}_1\right), ..., \sigma\left(\boldsymbol{w}^{(j)}\cdot\boldsymbol{z}_d\right)\right\} - \max\left\{\sigma\left(\boldsymbol{u}^{(j)}\cdot\boldsymbol{z}_1\right), ..., \sigma\left(\boldsymbol{u}^{(j)}\cdot\boldsymbol{z}_d\right)\right\}\right]$$

[6] It holds that $35c_\eta - \frac{1-\frac{5c_\eta}{4}}{\alpha(k)+1} < 0$ if and only if $\alpha(k) < \beta$ which holds if and only if $k > 64\left(\frac{\beta+1}{\beta-1}\right)^2$.

Let E denote the event for which at least one of the following holds:

1. $W_T^+(1) = \emptyset$.
2. $W_T^+(3) = \emptyset$.
3. $\boldsymbol{u}^{(1)} \cdot \boldsymbol{p}_2 > 0$ and $\boldsymbol{u}^{(2)} \cdot \boldsymbol{p}_2 > 0$.
4. $\boldsymbol{u}^{(1)} \cdot \boldsymbol{p}_4 > 0$ and $\boldsymbol{u}^{(2)} \cdot \boldsymbol{p}_4 > 0$.

Our proof will proceed as follows. We will first show that if E occurs then gradient descent does not learn f^*, i.e., the network N does not satisfy $\operatorname{sign}(N(\boldsymbol{x})) = f^*(\boldsymbol{x})$ for all $\boldsymbol{x} \in \{\pm 1\}^{2d}$. Then, we will show that $\mathbb{P}[E] \geq \frac{11}{12}$. This will conclude the proof.

Assume that one of the first two items in the definition of the event E occurs. Without loss of generality assume that $W_T^+(1) = \emptyset$ and recall that $\boldsymbol{x}^-$ denotes a negative vector which only contains the patterns $\boldsymbol{p}_2, \boldsymbol{p}_4$ and let $\boldsymbol{z}^+ \in \mathbb{R}^{2d}$ be a positive vector which only contains the patterns $\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_4$. By the assumption $W_T^+(1) = \emptyset$ and the fact that $\boldsymbol{p}_1 = -\boldsymbol{p}_3$ it follows that for all $j = 1, 2$,

$$\max\left\{\sigma\left(\boldsymbol{w}^{(j)} \cdot \boldsymbol{z}_1^+\right), ..., \sigma\left(\boldsymbol{w}^{(j)} \cdot \boldsymbol{z}_d^+\right)\right\} = \max\left\{\sigma\left(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{w}^{(j)} \cdot \boldsymbol{x}_d^-\right)\right\}$$

Furthermore, since $\boldsymbol{z}^+$ contains more distinct patterns than $\boldsymbol{x}^-$, it follows that for all $j = 1, 2$,

$$\max\left\{\sigma\left(\boldsymbol{u}^{(j)} \cdot \boldsymbol{z}_1^+\right), ..., \sigma\left(\boldsymbol{u}^{(j)} \cdot \boldsymbol{z}_d^+\right)\right\} \geq \max\left\{\sigma\left(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}_1^-\right), ..., \sigma\left(\boldsymbol{u}^{(j)} \cdot \boldsymbol{x}_d^-\right)\right\}$$

Hence, $N(\boldsymbol{z}^+) \leq N(\boldsymbol{x}^-)$. Since at a global minimum $N(\boldsymbol{x}^-) \leq -1$, we have $N(\boldsymbol{z}^+) \leq -1$ and $\boldsymbol{z}_2$ is not classified correctly.

Now assume without loss of generality that the third item in the definition of E occurs. Let $\boldsymbol{z}^-$ be the negative vector with all of its patterns equal to $\boldsymbol{p}_4$. It is clear that $N(\boldsymbol{z}^-) \geq 0$ and therefore $\boldsymbol{z}^-$ is not classified correctly. This concludes the first part of the proof. We will now proceed to show that $\mathbb{P}[E] \geq \frac{11}{12}$.

Denote by A_i the event that item i in the definition of E occurs and for an event A denote by A^c its complement. Thus $E^c = \cap_{i=1}^4 A_i^c$ and $\mathbb{P}[E^c] = \mathbb{P}[A_3^c \cap A_4^c \mid A_1^c \cap A_2^c]\,\mathbb{P}[A_1^c \cap A_2^c]$.

We will first calculate $\mathbb{P}[A_1^c \cap A_2^c]$. By Lemma E.5, we know that for $i \in \{1, 3\}$, $W_0^+(i) = W_T^+(i)$. Therefore, it suffices to calculate the probabilty that $W_0^+(1) \neq \emptyset$ and $W_0^+(3) \neq \emptyset$, provided that $W_0^+(1) \cup W_0^+(3) \neq \emptyset$. Without conditioning on $W_0^+(1) \cup W_0^+(3) \neq \emptyset$, for each $1 \leq i \leq 4$ and $1 \leq j \leq 2$ the event that $j \in W_0^+(i)$ holds with probability $\frac{1}{4}$. Since the initializations of the filters are independent, we have $\mathbb{P}[A_1^c \cap A_2^c] = \frac{1}{6}$. [7]

We will show that $\mathbb{P}[A_3^c \cap A_4^c \mid A_1^c \cap A_2^c] = \frac{1}{2}$ by a symmetry argument. This will finish the proof of the theorem. For the proof, it will be more convenient to denote the matrix of weights at iteration t as a tuple of 4 vectors, i.e., $W_t = \left(\boldsymbol{w}_0^{(1)}, \boldsymbol{w}_0^{(2)}, \boldsymbol{u}_0^{(1)}, \boldsymbol{u}_0^{(2)}\right)$. Consider two initializations $W_0^{(1)} = \left(\boldsymbol{w}_0^{(1)}, \boldsymbol{w}_0^{(2)}, \boldsymbol{u}_0^{(1)}, \boldsymbol{u}_0^{(2)}\right)$ and $W_0^{(2)} = \left(\boldsymbol{w}_0^{(1)}, \boldsymbol{w}_0^{(2)}, -\boldsymbol{u}_0^{(1)}, \boldsymbol{u}_0^{(2)}\right)$ and let $W_t^{(1)}$ and $W_t^{(2)}$ be the corresponding weight values at iteration t. We will prove the following lemma:

Lemma F.2. *For all $t \geq 0$, if $W_t^{(1)} = \left(\boldsymbol{w}_t^{(1)}, \boldsymbol{w}_t^{(2)}, \boldsymbol{u}_t^{(1)}, \boldsymbol{u}_t^{(2)}\right)$ then $W_t^{(2)} = \left(\boldsymbol{w}_t^{(1)}, \boldsymbol{w}_t^{(2)}, -\boldsymbol{u}_t^{(1)}, \boldsymbol{u}_t^{(2)}\right)$.*

Proof. We will show this by induction on t. [8]This holds by definition for $t = 0$. Assume it holds for an iteration t. Denote $W_{t+1}^{(2)} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{v}_1, \boldsymbol{v}_2)$. We need to show that $\boldsymbol{z}_1 = \boldsymbol{w}_{t+1}^{(1)}$, $\boldsymbol{z}_2 = \boldsymbol{w}_{t+1}^{(2)}$, $\boldsymbol{v}_1 = -\boldsymbol{u}_{t+1}^{(1)}$ and $\boldsymbol{v}_2 = \boldsymbol{u}_{t+1}^{(2)}$. By the induction hypothesis it holds that $N_{W_t^{(1)}}(\boldsymbol{x}^+) = N_{W_t^{(2)}}(\boldsymbol{x}^+)$ and $N_{W_t^{(1)}}(\boldsymbol{x}^-) = N_{W_t^{(2)}}(\boldsymbol{x}^-)$. This follows since for diverse points (either positive or negative), negating

[7]Note that this holds after conditioning on the corresponding event of Lemma E.3.

[8]Recall that we condition on the event corresponding to Lemma E.3. By negating a weight vector we still satisfy the bounds in the lemma and therefore the claim that will follow will hold under this conditioning.

a neuron does not change the function value. Thus, according to Eq. 2 and Eq. 3 we have $\boldsymbol{z}_1 = \boldsymbol{w}_{t+1}^{(1)}$, $\boldsymbol{z}_2 = \boldsymbol{w}_{t+1}^{(2)}$ and $\boldsymbol{v}_2 = \boldsymbol{u}_{t+1}^{(2)}$. We are left to show that $\boldsymbol{v}_1 = -\boldsymbol{u}_{t+1}^{(1)}$. This follows from Eq. 3 and the following facts:

1. $\boldsymbol{p}_3 = -\boldsymbol{p}_1$.
2. $\boldsymbol{p}_2 = -\boldsymbol{p}_4$.
3. $\arg\max_{1\leq l\leq 4} \boldsymbol{u}\cdot\boldsymbol{p}_l = 1$ if and only if $\arg\max_{1\leq l\leq 4} -\boldsymbol{u}\cdot\boldsymbol{p}_l = 3$.
4. $\arg\max_{1\leq l\leq 4} \boldsymbol{u}\cdot\boldsymbol{p}_l = 2$ if and only if $\arg\max_{1\leq l\leq 4} -\boldsymbol{u}\cdot\boldsymbol{p}_l = 4$.
5. $\arg\max_{l\in\{2,4\}} \boldsymbol{u}\cdot\boldsymbol{p}_l = 2$ if and only if $\arg\max_{l\in\{2,4\}} -\boldsymbol{u}\cdot\boldsymbol{p}_l = 4$.

To see this, we will illustrate this through one case, the other cases are similar. Assume, for example, that $\arg\max_{1\leq l\leq 4} \boldsymbol{u}_t^{(1)}\cdot\boldsymbol{p}_l = 3$ and $\arg\max_{l\in\{2,4\}} \boldsymbol{u}_t^{(1)}\cdot\boldsymbol{p}_l = 2$ and assume without loss of generality that $N_{W_t^{(1)}}(\boldsymbol{x}^+) = N_{W_t^{(2)}}(\boldsymbol{x}^+) < \gamma$ and $N_{W_t^{(1)}}(\boldsymbol{x}^-) = N_{W_t^{(2)}}(\boldsymbol{x}^-) > -1$. Then, by Eq. 3, $\boldsymbol{u}_{t+1}^{(1)} = \boldsymbol{u}_t^{(1)} - \boldsymbol{p}_3 + \boldsymbol{p}_2$. By the induction hypothesis and the above facts it follows that $\boldsymbol{v}_1 = -\boldsymbol{u}_t^{(1)} - \boldsymbol{p}_1 + \boldsymbol{p}_4 = -\boldsymbol{u}_t^{(1)} + \boldsymbol{p}_3 - \boldsymbol{p}_2 = -\boldsymbol{u}_{t+1}^{(1)}$. This concludes the proof. □

Consider an initialization of gradient descent where $\boldsymbol{w}_0^{(1)}$ and $\boldsymbol{w}_0^{(2)}$ are fixed and the event that we conditioned on in the beginning of the proof and $A_1^c \cap A_2^c$ hold. Define the set B_1 to be the set of all pair of vectors $(\boldsymbol{v}_1, \boldsymbol{v}_2)$ such that if $\boldsymbol{u}_0^{(1)} = \boldsymbol{v}_1$ and $\boldsymbol{u}_0^{(1)} = \boldsymbol{v}_2$ then at iteration T, $\boldsymbol{u}^{(1)}\cdot\boldsymbol{p}_2 > 0$ and $\boldsymbol{u}^{(2)}\cdot\boldsymbol{p}_2 > 0$. Note that this definition implicitly implies that this initialization satisfies the condition in Lemma E.3 and leads to a global minimum. Similarly, let B_2 be the set of all pair of vectors $(\boldsymbol{v}_1, \boldsymbol{v}_2)$ such that if $\boldsymbol{u}_0^{(1)} = \boldsymbol{v}_1$ and $\boldsymbol{u}_0^{(1)} = \boldsymbol{v}_2$ then at iteration T, $\boldsymbol{u}^{(1)}\cdot\boldsymbol{p}_4 > 0$ and $\boldsymbol{u}^{(2)}\cdot\boldsymbol{p}_2 > 0$. First, if $(\boldsymbol{v}_1, \boldsymbol{v}_2) \in B_1$ then $(-\boldsymbol{v}_1, \boldsymbol{v}_2)$ satisfies the conditions of Lemma E.3. Second, by Lemma F.2, it follows that if $(\boldsymbol{v}_1, \boldsymbol{v}_2) \in B_1$ then initializating with $(-\boldsymbol{v}_1, \boldsymbol{v}_2)$, leads to the same values of $N_{W_t}(\boldsymbol{x}^+)$ and $N_{W_t}(\boldsymbol{x}^-)$ in all iterations $0 \leq t \leq T$. Therefore, initializing with $(-\boldsymbol{v}_1, \boldsymbol{v}_2)$ leads to a convergence to a global minimum with the same value of T as the initialization with $(\boldsymbol{v}_1, \boldsymbol{v}_2)$. Furthermore, if $(\boldsymbol{v}_1, \boldsymbol{v}_2) \in B_1$, then by Lemma F.2, initializing with $\boldsymbol{u}_0^{(1)} = -\boldsymbol{v}_1$ and $\boldsymbol{u}_0^{(1)} = \boldsymbol{v}_2$ results in $\boldsymbol{u}^{(1)}\cdot\boldsymbol{p}_2 < 0$ and $\boldsymbol{u}^{(2)}\cdot\boldsymbol{p}_2 > 0$. It follows that $(\boldsymbol{v}_1, \boldsymbol{v}_2) \in B_1$ if and only if $(-\boldsymbol{v}_1, \boldsymbol{v}_2) \in B_2$.

For $l_1, l_2 \in \{2, 4\}$ define $P_{l_1,l_2} = \mathbb{P}\left[\boldsymbol{u}^{(1)}\cdot\boldsymbol{p}_{l_1} > 0 \wedge \boldsymbol{u}^{(2)}\cdot\boldsymbol{p}_{l_2} > 0 \mid A_1^c \cap A_2^c, \boldsymbol{w}_0^{(1)}, \boldsymbol{w}_0^{(2)}\right]$ Then, by symmetry of the initialization and the latter arguments it follows that $P_{2,2} = P_{4,2}$.

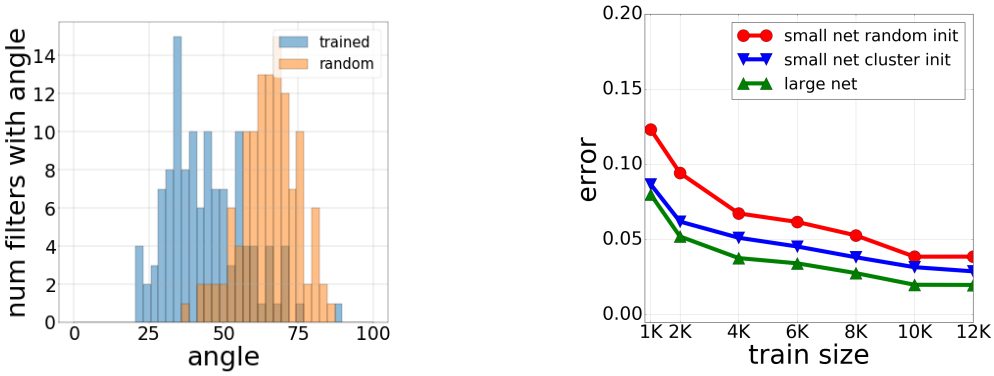
By similar arguments we can obtain the equalities $P_{2,2} = P_{4,2} = P_{4,4} = P_{2,4}$.

Since all of these four probabilities sum to 1, each is equal to $\frac{1}{4}$. [9]Taking expectations of these probabilities with respect to the values of $\boldsymbol{w}_0^{(1)}$ and $\boldsymbol{w}_0^{(2)}$ (given that Lemma E.3 and $A_1^c \cap A_2^c$ hold) and using the law of total expectation, we conclude that

$$\begin{aligned}\mathbb{P}\left[A_3^c \cap A_4^c \mid A_1^c \cap A_2^c\right] &= \mathbb{P}\left[\boldsymbol{u}^{(1)}\cdot\boldsymbol{p}_4 > 0 \wedge \boldsymbol{u}^{(2)}\cdot\boldsymbol{p}_2 > 0 \mid A_1^c \cap A_2^c\right] \\ &\quad + \mathbb{P}\left[\boldsymbol{u}^{(1)}\cdot\boldsymbol{p}_2 > 0 \wedge \boldsymbol{u}^{(2)}\cdot\boldsymbol{p}_4 > 0 \mid A_1^c \cap A_2^c\right] = \frac{1}{2}\end{aligned}$$

Finally, let $\mathcal{Z}_1$ be the set of positive points which contain only the patterns $\boldsymbol{p}_1$, $\boldsymbol{p}_2$, $\boldsymbol{p}_4$, $\mathcal{Z}_2$ be the set of positive points which contain only the patterns $\boldsymbol{p}_3$, $\boldsymbol{p}_2$, $\boldsymbol{p}_4$. Let $\mathcal{Z}_3$ be the set which contains the negative point with all patterns equal to $\boldsymbol{p}_2$ and $\mathcal{Z}_4$ be the set which contains the negative point with all patterns equal to $\boldsymbol{p}_4$. By the proof of the previous section, if the event E holds, then there exists $1 \leq i \leq 4$, such that gradient descent converges to a solution at iteration T which errs on all of the points in $\mathcal{Z}_i$. Therefore, its test error will be at least p^* (recall Eq. 5).

[9]Note that the probablity that $\boldsymbol{u}^{(i)}\cdot\boldsymbol{p}_j = 0$ is 0 for all possible i and j.



(a) (b)

Figure 3: Clustering and Exploration in MNIST with 4x4 filters (a) Distribution of angle to closest center in trained and random networks. (b) The plot shows the test error of the small network (4 channels) with standard training (red), the small network that uses clusters from the large network (blue), and the large network (120 channels) with standard training (green).

G Proof of Theorem 6.5

Let $\delta \geq 1 - p_+p_-(1 - c - 16e^{-8})$. By Theorem 6.3, given 2 samples, one positive and one negative, with probability at least $1 - \delta \leq p_+p_-(1 - c - 16e^{-8})$, gradient descent will converge to a global minimum that has 0 test error. Therefore, for all $\epsilon \geq 0$, $m(\epsilon, \delta) \leq 2$. On the other hand, by Theorem 6.4, if $m < \frac{2\log\left(\frac{48\delta}{33(1-c)}\right)}{\log(p_+p_-)}$ then with probability greater than

$$(p_+p_-)^{\frac{\log\left(\frac{48\delta}{33(1-c)}\right)}{\log(p_+p_-)}}(1-c)\frac{33}{48} = \delta$$

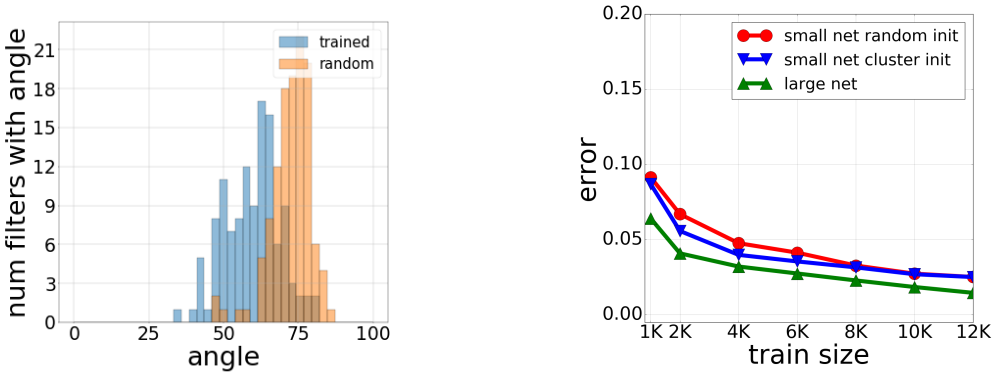
gradient descent converges to a global minimum with test error at least p^*. It follows that for $0 \leq \epsilon < p^*$, $m(\epsilon, \delta) \geq \frac{2\log\left(\frac{48\delta}{33(1-c)}\right)}{\log(p_+p_-)}$.

H Experiments for Section 7

We first provide several details on the experiments in Section 7. We trained the overparamaterized network with 120 channels once for each training set size and recorded the clustered weights. We used Adam for optimization and batch size which is one-tenth of the size of the training set. We used learning rate=0.01 and standard deviation of 0.05 for initialization with truncated normal weights. For the small network with random initialization we used the same optimization method and batch sizes but tried 6 different pairs of values for learning rate and standard deviation: (0.01,0.01), (0.01,0.05), (0.05,0.05), (0.05, 0.01), (0.1,0.5) and (0.1,0.1). For each pair and training set size we trained 20 times and averaged the results. The curve is the best test accuracy we got among all learning rate and standard deviation pairs.

For the small network with cluster initialization we experimented with the same setup as the small network with random initializatoin but only experimented with learning rate 0.01 and standard deviation 0.05. The curve is an average of 20 runs for each training set size.

We also experimented with other filter sizes in similar setups. Figure 3 shows the results for 4x4 filters and clustering from 120 filters to 4 filters (with 2000 training points). Figure 4 shows the results for 7x7 filters and clustering from 120 filters to 4 filters (with 2000 training points).



(a) (b)

Figure 4: Clustering and Exploration in MNIST with 7x7 filters (a) Distribution of angle to closest center in trained and random networks. (b) The plot shows the test error of the small network (4 channels) with standard training (red), the small network that uses clusters from the large network (blue), and the large network (120 channels) with standard training (green).

References

Vershynin, R. High-dimensional probability. *An Introduction with Applications*, 2017.