Contents

A Corrections to the main paper

In the course of preparing the supplementary materials we identified the following two mistakes.

1. First we found a mistake in one of the proofs for a lemma describing the conditions at initialization. We have fixed this issue in the supplementary and highlight below the two errors this has caused in the main paper.
   - An entry in the Table 1 needs to be updated to reflect these changes: in particular the dependence on n is $\frac{1}{\delta}$ not $n \geq 1$ for Theorem 3.1. Aside from this one entry Table 1 is correct. For the convenience of the reader we provide the full, corrected table below.

Table 1: Comparison of results up to constants, note in all cases $d \geq Cn^2 \log(n/\delta)$, $k \leq \frac{n}{C}$ where C is an appropriatly chosen constant.

	Frei et al. (2022)	Xu & Gu (2023)	Theorem 3.1	Theorem 3.6	Theorem 3.8
$n \geq C\cdot$	$\log\left(\frac{1}{\delta}\right)$	$\log\left(\frac{m}{\delta}\right)$	$\frac{1}{\delta}$	1	$\log\left(\frac{m}{\delta}\right)$
$m \geq C\cdot$	1	$\log\left(\frac{n}{\delta}\right)$	$\log\left(\frac{n}{\delta}\right)$	$\log\left(\frac{n}{\delta}\right)$	$\log\left(\frac{n}{\delta}\right)$
$\gamma \leq \frac{1}{C}\cdot$	$\frac{1}{n}$	$\frac{1}{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{nd}}$	$\frac{1}{k}$
$\gamma \geq C\cdot$	$\frac{1}{\sqrt{nd}}$	$\sqrt{\frac{\log(\frac{md}{n\delta})}{nd}}$	$\sqrt{\frac{\log(\frac{n^2}{\delta})}{d}}$	0	$\frac{1}{n}$
Result	Benign[1]	Benign	Benign	Non-benign	No overfit

   - The same mistake also means that the sentence starting on line 188 “Comparing specifically the results on benign overfitting, we observe a better dependency on n in particular compared with Xu & Gu (2023)..." is incorrect and indeed their work has better dependence on n.
2. On line 262 we made a mistake in the explanation of our results: instead of $C\epsilon(1 + n\gamma)$ it should read $\frac{C\varepsilon n\gamma}{(\eta(\gamma+\rho)}$.

B Problem setup

B.1 Notation

In order to provide a convenient reference for the reader, we summarize our notation as follows.

- For $n \in \mathbb{Z}_{\geq 1}$ then $[n] := \{1, 2, 3...n\}$. Furthermore we let $[n]_e := \{i \in [n] : \ i \text{ is even}\}$ and $[n]_o := \{i \in [n] : \ i \text{ is odd}\}$.
- For two iterations t_0, t_1 with $t_1 > t_0$ we define the following counting functions.
  1. $T_{ij}(t_0, t_1) := \sum_{\tau=t_0}^{t_1-1} \mathbb{1}(i \in \mathcal{A}_j^{(\tau)} \cap \mathcal{F}^{(\tau)})$ is the number of times the i-th data point updates the j-th neuron between iterations t_0 and t_1.
  2. $T_i(t_0, t_1) := \sum_{j \in [2m]} T_{ij}(t_0, t_1)$ is the total number of updates from the i-th data point to the entire network between iterations t_0 and t_1.
  3. $G_j(t_0, t_1) := \sum_{i \in \mathcal{S}_T} T_{ij}(t_0, t_1)$ and $B_j(t_0, t_1) := \sum_{i \in \mathcal{S}_F} T_{ij}(t_0, t_1)$ are the number of clean and corrupt updates applied to the j-th neuron respectively between iterations t_0 and t_1. We further define $G_j^{(i)}(t_0, t_1) := G_j(t_0, t_1) - \mathbb{1}(i \in \mathcal{S}_T) T_{ij}(t_0, t_1)$ and $B_j^{(i)}(t_0, t_1) := B_j(t_0, t_1) - \mathbb{1}(i \in \mathcal{S}_F) T_{ij}(t_0, t_1)$.
  4. $H_j^{(i)}(t_0, t_1) := \sum_{\ell \neq i} T_{\ell j}(t_0, t_1)$ is the number of times any data point except the i-th updates the j-th neuron between iterations t_0 and t_1.
  5. $G(t_0, t_1) := \sum_{j \in [2m]} G_j(t_0, t_1)$ and $B(t_0, t_1) := \sum_{j \in [2m]} B_j(t_0, t_1)$ are the total number of clean and corrupt updates applied to the entire network between iterations t_0 and t_1. We similarly define $G^{(i)}(t_0, t_1) = G(t_0, t_1) - \mathbb{1}(i \in \mathcal{S}_T) T_i(t_0, t_1)$ and $B^{(i)}(t_0, t_1) = B(t_0, t_1) - \mathbb{1}(i \in \mathcal{S}_F) T_i(t_0, t_1)$
  6. $T(t_0, t_1) := \sum_{\ell} T_\ell(t_0, t_1)$ is the total number of updates from all points applied to the entire network between iterations t_0 and t_1. We also define $T^{(i)}(t_0, t_1) = T(t_0, t_1) - T_i(t_0, t_1)$, the number of updates excluding those from point i.
  7. $S_i(t_0, t_1) := \sum_{\tau=t_0}^{t_1-1} \mathbb{1}(\exists j \in [2m] : \ i \in \mathcal{A}_j^{(\tau)} \cap \mathcal{F}^{(\tau)})$ is the number of iterations between t_0 and t_1 in which the ith data point participates. We say a data point participates during an iteration if it contributes to the update of at least one neuron at said iteration.
- We extend each of these definitions to the case $t_0 = t_1$ by letting the empty sum be zero.
- The signal alignment of the j-th neuron is defined as $C_j^{(t)} = \langle \mathbf{w}_j^{(t)}, (-1)^j \mathbf{v} \rangle$.
- We use the notation $O(\eta)$ to denote a quantity $f(\eta)$ such that

$$\limsup_{\eta \to 0} \frac{|f(\eta)|}{\eta} < \infty.$$

  Likewise, $f(\eta) = O(1)$ if

$$\limsup_{\eta \to 0} |f(\eta)| < \infty,$$

  $f(\eta) = \Omega(\eta)$ if

$$\liminf_{\eta \to 0} \frac{|f(\eta)|}{\eta} > 0,$$

  and $f(\eta) = \Omega(1)$ if

$$\liminf_{\eta \to 0} |f(\eta)| > 0.$$

  Here the limit is taken as η and λ_w go to zero while the other parameters of the model and data remain fixed. We will always choose λ_w such that $\lambda_w = O(\eta)$.
- Denote $\mathcal{T}_0$ be the first iteration t where $\mathcal{F}^{(t)} \neq [2n]$.
- We use $C \geq 1$ and $c \leq 1$ to generically represent sufficiently large and small constants respectively. Furthermore, we reuse both C and c from one line to another: for example, $2Cx = Cx$ and $0.5cx = cx$.

- Finally, in much of our analysis for particular variables, notably γ, the constants involved matter and as such we work with explicit constants. For other variables, in particular m and n, then typically as long as they are sufficiently large the explicit constants involved are not important. As such we typically resort to using a generically large enough constant C.

B.2 Data model

For the reader's convenience we recap the data model studied in this work. We consider a training sample of $2n$ pairs of points and their labels $(\mathbf{x}_i, y_i)_{i=1}^{2n}$ where $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{-1, +1\}$. Furthermore, we identify two disjoint subsets $\mathcal{S}_T \subset [2n]$ and $\mathcal{S}_F \subset [2n]$, where $\mathcal{S}_T \cup \mathcal{S}_F = [2n]$, which correspond to the clean and corrupt points in our sample respectively. The categorization of a point as clean or corrupted is determined by its label: for all $i \in [2n]$ we assume $y_i = \beta(i)(-1)^i$ where $\beta(i) = -1$ iff $i \in \mathcal{S}_F$ and $\beta(i) = 1$ otherwise. In addition, we assume $|\mathcal{S}_F \cap [2n]_e| = |\mathcal{S}_F \cap [2n]_o| = k$ and $|\mathcal{S}_T \cap [2n]_e| = |\mathcal{S}_T \cap [2n]_o| = n - k$. We remark that this assumption simplifies the exposition of our results but is not actually integral to our analysis. Each data point is assumed to have the form

$$\mathbf{x}_i = (-1)^i(\sqrt{\gamma}\mathbf{v} + \sqrt{1-\gamma}\beta(i)\mathbf{n}_i). \tag{1}$$

Here $\mathbf{v} \in \mathbb{R}^d$ satisfies $\|\mathbf{v}\| = 1$. We refer to $\mathbf{v}$ as the signal component as the alignment of a clean point with $\mathbf{v}$ determines its sign. Indeed, $\text{sign}(\langle \mathbf{x}_i, \mathbf{v}\rangle) = (-1)^i = y_i$ for $i \in \mathcal{S}_T$ whereas $\text{sign}(\langle \mathbf{x}_i, \mathbf{v}\rangle) = -y_i$ for $i \in \mathcal{S}_F$. Thus we may view the labels of corrupt point as flipped from their clean state. The random vectors $(\mathbf{n}_i)_{i=1}^{2n}$ are mutually independent and identically distributed (i.i.d.) drawn from the uniform distribution over $\mathbb{S}^{d-1} \cap \text{span}\{\mathbf{v}\}^\perp$, which we denote $U(\mathbb{S}^{d-1} \cap \text{span}\{\mathbf{v}\}^\perp)$. This distribution is symmetric, mean zero and for any $\mathbf{n} \sim U(\mathbb{S}^{d-1} \cap \text{span}\{\mathbf{v}\}^\perp)$ it holds that $\mathbf{n} \perp \mathbf{v}$ and $\|\mathbf{n}\| = 1$. We refer to these vectors as noise components due to the fact that they are independent of the labels of their respective points. We also remark that as the noise distribution is symmetric then clean and corrupt points are identically distributed. Indeed, the multiplication of the noise component by $\beta(i)$ results in the following expression which will prove convenient.

$$y_i\mathbf{x}_i = \beta(i)\sqrt{\gamma}\mathbf{v} + \sqrt{1-\gamma}\mathbf{n}_i. \tag{2}$$

This expression entails that the only difference between clean and corrupt points during training is that they push neurons in opposite directions along the signal vector. Finally, the real, scalar quantity $\gamma \in [0, 1]$ controls the strength of the signal versus noise, furthermore the clean margin, i.e., the distance from any clean point to the max margin classifier, is $\sqrt{\gamma}$ by construction. Thus far we have discussed only the data in the training sample. We assume test data are drawn mutually i.i.d. from the same distribution as the points in the training sample but with the added proviso that they are always clean: to be clear, at test time a label is sampled as $y \sim U(\{-1, 1\})$ and the corresponding data point has the form

$$\mathbf{x} = y(\sqrt{\gamma}\mathbf{v} + \sqrt{1-\gamma}\mathbf{n}), \tag{3}$$

where again $\mathbf{n} \sim U(\mathbb{S}^{d-1} \cap \text{span}\{\mathbf{v}\}^\perp)$.

B.3 Network architecture, optimization and initialization

Here we also recap the network architecture and optimization and initialization setting. We consider a densely connected, single layer feedforward neural network $f : \mathbb{R}^{2m\times d} \times \mathbb{R}^d \to \mathbb{R}^d$ with the following forward pass map,

$$f(\mathbf{W}, \mathbf{x}) = \sum_{j=1}^{2m} (-1)^j \phi(\langle \mathbf{w}_j, \mathbf{x}\rangle).$$

Here $\phi := \max\{0, z\}$ denotes the ReLU activation function, $\mathbf{w}_j$ the jth row of $\mathbf{W}$ and w_{jc} the element of $\mathbf{W}$ on row $j \in [m]$ and column $c \in [d]$. We remark that only the weights of the hidden layer, which we also refer to as the weights of the network, are trainable and the outer weights remain frozen throughout training. The network weights are optimized using full batch gradient descent (GD) with step size $\eta > 0$ to minimize the hinge loss over a training sample $((\mathbf{x}_i, y_i))_{i=1}^{2n} \subset (\mathbb{R}^d \times \{-1, 1\})^{2n}$ as described in Section B.2. After $t' > 0$ iterations this optimization process generates a sequence of weight matrices $(\mathbf{W}^{(t)})_{t=0}^{t'}$. For convenience we overload our

notation for the forward pass map of the network by letting $f(t, \mathbf{x}) := f(\mathbf{W}^{(t)}, \mathbf{x})$. We denote the hinge loss on the i-th point at iteration t as

$$\ell(t, i) := \max\{0, 1 - y_i f(t, \mathbf{x}_i)\},$$

the hinge loss over the entire training sample at iteration t is therefore defined as

$$L(t) := \sum_{i=1}^{2n} \ell(t, i).$$

Let $\mathcal{F}^{(t)} := \{i \in [2n] : \ell(t, \mathbf{x}_i) < 1\}$ be the set of points that have nonzero loss at iteration t, and $\mathcal{A}_j^{(t)} := \{i \in [2n] : \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle > 0\}$ the set of points which activate the jth neuron at iteration t. Combining (2) with

$$\frac{\partial \ell(t, i)}{\partial w_{jr}} = \begin{cases} 0, & \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle \leq 0, \\ -(-1)^j y_i x_{ir}, & \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle > 0 \end{cases}$$

gives that the GD update rule[2] for the neuron weights for any iteration $t \geq 0$,

$$\mathbf{w}_j^{(t+1)} = \mathbf{w}_j^{(t)} + (-1)^j \eta \sum_{\ell=1}^{2n} \mathbb{1}(\ell \in \mathcal{A}_j^{(t)} \cap \mathcal{F}^{(t)}) y_\ell \mathbf{x}_\ell. \tag{4}$$

In regard to the initialization of the network weights, for convenience we assume each neuron's weight vector is drawn mutually i.i.d. uniform from the centered sphere with radius $\lambda_w > 0$. We remark that results analogous to the ones presented trivially hold if the weights are instead initialized mutually i.i.d. as $w_{jc}^{(0)} \sim \mathcal{N}(0, \sigma_w^2)$ as long as σ_w^2 is sufficiently small.

B.4 Properties of the data and network at initialization

For each of our results to hold we require certain properties on both the network weights and training sample to hold at initialization. Here we bound the probabilities of each of these events in isolation and later will combine them using the union bound.

First, and in order to prove convergence, we require the noise components of the training sample to be approximately orthogonal to one another. A training sample whose noise components satisfy this approximate orthogonality condition we refer to as "good".

Lemma B.1. *Let $\rho, \delta \in (0, 1)$. Given a sequence $(\mathbf{n}_i)_{i=1}^{2n}$ of mutually i.i.d. random vectors with $\mathbf{n}_i \sim U(\mathbb{S}^{d-1} \cap \text{span}(\mathbf{v})^\perp)$, then assuming $d \geq \max\left\{6, 3\rho^{-2} \ln\left(\frac{2n^2}{\delta}\right)\right\}$*

$$\mathbb{P}\left(\bigcap_{i \neq \ell} \{|\langle \mathbf{n}_i, \mathbf{n}_\ell \rangle| \leq \rho\}\right) \geq 1 - \delta.$$

Proof. For any pairs of mutually i.i.d. random vectors $\mathbf{n}, \mathbf{n}' \sim U(\mathbb{S}^{d-1} \cap \text{span}(\mathbf{v})^\perp)$ and $\mathbf{u}, \mathbf{u}' \sim U(\mathbb{S}^{d-2})$ observe

$$\langle \mathbf{n}, \mathbf{n}' \rangle \overset{d}{=} \langle \mathbf{u}, \mathbf{u}' \rangle.$$

Due to independence of $\mathbf{u}, \mathbf{u}'$ and the rotational invariance of $U(\mathbb{S}^{d-2})$

$$\langle \mathbf{u}, \mathbf{u}' \rangle \overset{d}{=} \langle \mathbf{u}, \mathbf{e}_1 \rangle,$$

where here $\mathbf{e}_1 = [1, 0..0]^T$. Let $\text{Cap}(\mathbf{e}_1, \rho) := \{\mathbf{z} \in \mathbb{S}^{d-2} : \langle \mathbf{e}_1, \mathbf{z} \rangle \geq \rho\}$ denote the *spherical cap* of $\mathbb{S}^{d-2}$ centered on $\mathbf{e}_1$. As $d \geq 6$ then from Ball (1997)[Lemma 2.2] it follows that

$$\mathbb{P}(|\langle \mathbf{n}, \mathbf{n}' \rangle| \geq \rho) = \mathbb{P}(\mathbf{u} \in \text{Cap}(\mathbf{e}_1, \rho)) \leq \exp\left(-\frac{(d-2)\rho^2}{2}\right) \leq \exp\left(-\frac{d\rho^2}{3}\right).$$

[2] Although the derivative of ReLU clearly does not exist at zero, we follow the routine procedure of defining an update rule that extends the gradient update to cover this event.

Applying the union bound then

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i,\ell\in[2n],i\neq\ell}\{|\langle\mathbf{n}_i,\mathbf{n}_\ell\rangle|\leq\rho\}\right) &= 1-\mathbb{P}\left(\bigcup_{i,\ell\in[2n],i\neq\ell}\{|\langle\mathbf{n}_i,\mathbf{n}_\ell\rangle|\geq\rho\}\right)\\ &\geq 1-2n^2\mathbb{P}\left(|\langle\mathbf{n}_i,\mathbf{n}_\ell\rangle|\geq\rho\right)\\ &\geq 1-2n^2\exp\left(-\frac{d\rho^2}{3}\right).\end{aligned}$$

Setting $\delta \geq 2n^2\exp\left(-\frac{d\rho^2}{3}\right)$ and rearranging we arrive at the result claimed. □

In addition to requiring the approximate orthogonality property on the training data, our approach also requires at least a large proportion of the neurons at initialization to satisfy certain conditions. To this end, we introduce the following notation, where $p \in \{-1, 1\}$.

- Let $\Gamma_p := \{j \ : \ (-1)^j = p,\ G_j(0,1)(\gamma-\rho) - B_j(0,1)(\gamma+\rho) \geq \frac{2\lambda_w}{\eta}\}$ denote the set of neurons with output weight $(-1)^p$ which have more clean points activating them than corrupt ones at initialization. We will show that these sets of neurons have a *predictable* behavior early during training before any clean points achieve zero loss. Let $\Gamma = \Gamma_1 \cup \Gamma_{-1}$.
- Let $\Theta_p := \{j \sim \Gamma_p \ : \ G_j(0,1)(\gamma+\rho) - B_j(0,1)(\gamma-\rho) < 1-\gamma+\rho\} \subset \Gamma_p$. We will show that neurons in this subset are able to *carry* corrupt points through training, eventually, at least in the overfitting setting, enabling them to achieve zero loss.. Let $\Theta = \Theta_1 \cup \Theta_{-1}$.

Our goals are two-fold: first show Γ_p accounts for a significant proportion of the neurons with output label sign matching p, second, and of particular importance for our result on benign overfitting, ensure each corrupt point activates a neuron in Θ_p where p matches its label. To this end we first provide the following result.

Lemma B.2. *Define $\mu := \frac{2k}{n+k}$ and assume $\kappa \in (0,1)$ satisfies $\kappa > \mu$. Given an arbitrary neuron $\mathbf{w}_j \sim U(\mathbb{S}^{d-1})$, we say that a collection of training points is (ε,κ)-good iff both $T_j(0,1) \geq 1$ and $B_j(0,1) < \kappa T_j(0,1)$ with probability at least $1-\epsilon$ over the randomness of the neuron. Define*

$$\delta := 2\exp\left(-\frac{(n+k)(\kappa-\mu)^2}{16}\right),$$

then with probability at least $1-\frac{\delta}{\epsilon}$ the training sample is (ε,κ)-good.

Proof. First we establish the notation for what follows: we say a point $\mathbf{x}$ is positive iff $\langle\mathbf{x},\mathbf{v}\rangle > 0$ and is negative iff $\langle\mathbf{x},\mathbf{v}\rangle < 0$. We use $\mathcal{S}^+$ and $\mathcal{S}^-$ to denote these sets of points respectively. Note by construction, see (1), clean and corrupt points of the same sign are mutually i.i.d. As here we only ever consider the activations at initialization, we also drop both the subscript j as well as the argument parentheses on the counting functions. We also use $\pm$ superscripts to denote the subsets corresponding to activations from positive and negative points respectively: as indicative examples of this notation, T is therefore used as shorthand for the total number of activations, B^+ is the number corrupt positive activations and G^- is the number of clean negative activations.

By the symmetry of the distribution of $\mathbf{w}$, $\mathbb{P}(\langle\mathbf{w},\mathbf{v}\rangle > 0) = \mathbb{P}(\langle\mathbf{w},\mathbf{v}\rangle < 0) = \frac{1}{2}$. As a result

$$\begin{aligned}\mathbb{P}((B<\kappa T)\cap(T>0)) = &\frac{1}{2}\mathbb{P}((B<\kappa T)\cap(T>0)\mid\langle\mathbf{w},\mathbf{v}\rangle>0)\\ &+\frac{1}{2}\mathbb{P}((B<\kappa T)\cap(T>0)\mid\langle\mathbf{w},\mathbf{v}\rangle<0).\end{aligned}$$

As the analysis and results derived under either condition will prove identical under reversal of the signs involved, without loss of generality we let $\langle\mathbf{w},\mathbf{v}\rangle > 0$. Using the union bound

$$\mathbb{P}((B<\kappa T)\cap(T>0)\mid\langle\mathbf{w},\mathbf{v}\rangle>0) \geq 1-\mathbb{P}(T=0\mid\langle\mathbf{w},\mathbf{v}\rangle>0)-\mathbb{P}(B\geq\kappa T\mid\langle\mathbf{w},\mathbf{v}\rangle>0),$$

therefore it suffices to upper bound the two probabilities on the right-hand-side.

Observe if $\langle \mathbf{w}, \mathbf{v} \rangle > 0$ then for $\mathbf{x} \in \mathcal{S}^+$ $\mathbb{P}(\langle \mathbf{x}, \mathbf{w} \rangle) > 1/2$ and for $\mathbf{x} \in \mathcal{S}^-$ $\mathbb{P}(\langle \mathbf{x}, \mathbf{w} \rangle) < 1/2$. By the mutual independence of the preactivations $(\langle \mathbf{x}, \mathbf{w}_j \rangle)_{i=1}^{2n}$ then $\mathbb{P}(T = 0 \mid \langle \mathbf{w}, \mathbf{v} \rangle > 0) \leq (1/2)^n$. To upper bound the other probability of interest we further condition on the following two events: there are no negative clean activations and all corrupt points are positive. Conditioning on these three events, which we denote for convenience Λ, then $T^+ = T$ and furthermore the event $B < \kappa T$ is equivalent to $B^+ < \kappa T^+$. Again as the preactivations are mutually independent the number of positive activations can be lower bounded using a binomial distribution with probability $1/2$. Applying a Chernoff bound

$$\mathbb{P}\left(T^+ \geq \frac{n+k}{4} \;\middle|\; \Lambda\right) \geq 1 - \exp\left(-\frac{n+k}{16}\right).$$

Furthermore, sampling positive points which activate $\mathbf{w}_j$ is equivalent to uniformly sampling without replacement T^+ points from $\mathcal{S}^+$. Let $Z_\ell = 1$ iff the ℓ-th element sampled from $\mathcal{S}^+$ is corrupt and is 0 otherwise. Using a variant of Hoeffding's bound for sampling without replacement (see Proposition 1.2 of Bardenet & Maillard (2015) for example)

$$\mathbb{P}\left(B^+ \geq \kappa T^+ \mid \Lambda\right) = \mathbb{P}\left(\frac{1}{T^+}\sum_{\ell=1}^{T^+} Z_\ell - \mu \geq \kappa - \mu\right) \leq \exp\left(-2T^+(\kappa-\mu)^2\right).$$

Therefore

$$\begin{aligned}
&\mathbb{P}((B < \kappa T) \cap (T > 0) \mid \langle \mathbf{w}, \mathbf{v} \rangle > 0) \\
&\geq 1 - \mathbb{P}(T = 0 \mid \langle \mathbf{w}, \mathbf{v} \rangle > 0) - \mathbb{P}(B \geq \kappa T \mid \langle \mathbf{w}, \mathbf{v} \rangle > 0) \\
&\geq 1 - (1/2)^n - \mathbb{P}\left(B^+ \geq \kappa T^+ \;\middle|\; T^+ \geq \frac{n+k}{4}, \Lambda\right)\mathbb{P}\left(T^+ \geq \frac{n+k}{4} \;\middle|\; \Lambda\right) \\
&\geq 1 - (1/2)^n - \exp\left(-\frac{(n+k)(\kappa-\mu)^2}{16}\right) \\
&\geq 1 - \delta.
\end{aligned}$$

Now if instead $\langle \mathbf{x}, \mathbf{w} \rangle < 0$ then swapping the roles of the negative and positive points in the argument above gives the same answer. As a result

$$\mathbb{P}((B < \kappa T) \cap (T > 0)) \geq 1 - \delta.$$

For convenience let $X := (\mathbf{x}_i)_{i=1}^{2n}$ and

$$\chi^c_{\kappa,\epsilon} = \{X : \mathbb{P}_w((B \geq \kappa T) \cup (T = 0)) > \epsilon\}.$$

Note here that the subscript w indicates randomness over the neuron alone and in addition clearly by construction

$$\mathbb{P}\left((B \geq \kappa T) \cup (T = 0) \mid X \in X^c_{\kappa,\epsilon}\right) > \epsilon.$$

Furthermore, as

$$\delta \geq \mathbb{P}\left((B \geq \kappa T) \cup (T = 0)\right) \geq \mathbb{P}\left((B \geq \kappa T) \cup (T = 0) \mid X \in X^c_{\kappa,\epsilon}\right)\mathbb{P}(X \in \chi^c_{\kappa,\epsilon}),$$

then it follows that $\mathbb{P}\left(X \in X^c_{\kappa,\epsilon}\right) \leq \frac{\delta}{\epsilon}$. As a result we conclude that the probability of drawing a (κ, ϵ)-good training sample is at least $1 - \frac{\delta}{\epsilon}$. □

Based on Lemma B.2, the following lemma bounds the probability that Γ_p is sufficiently large for our purposes. In particular, for our result on non-overfitting we require $|\Gamma_p| = m$, while for our benign overfitting result only that $|\Gamma_p| \geq (1-\alpha)m$ for some constant $\alpha \in (0, 1)$.

Lemma B.3. *Suppose $n \geq 15k$, $2\lambda_w \leq \eta(\gamma - \rho)$, $\gamma \geq 2\rho$ and $p \in \{-1, 1\}$. Then for sufficiently large n there exists a constant $c > 0$ such that the following are true.*

1. $\mathbb{P}\left(|\Gamma_p| = m\right) = 1 - m\exp(-cn)$.
2. *With $\alpha \in (0, 1)$ a constant such that $\alpha m \in [m]$, then*

$$\mathbb{P}\left(|\Gamma_p| \geq (1-\alpha)m\right) \geq 1 - \exp(-cn).$$

Proof. As here we only ever consider the activations at initialization, for convenience we drop the argument parentheses "$(0,1)$" on the counting functions: in particular we write $T_j(0,1)$ as T_j and $B_j(0,1)$ as B_j. Suppose $(j)^{-1} = p$, under the assumption $\lambda_w \leq \eta(\gamma - \rho)$ then if

$$G_j(\gamma - \rho) - B_j(\gamma + \rho) > (\gamma - \rho) \geq \frac{2\lambda_w}{\eta}$$

we may conclude $j \in \Gamma_p$. Rearranging this expression, equivalently $j \in \Gamma_p$ if

$$(1 + B_j)\gamma < (\gamma - \rho)T_j.$$

As a result, membership to Γ_p is guaranteed as long as $T_j > 0$ and $B_j < \frac{\gamma-\rho}{2\gamma}T_j$. Note by the assumptions of the lemma $\mu := \frac{2k}{n+k} \leq \frac{1}{8}$ and $\frac{\gamma-\rho}{2\gamma} \geq \frac{1}{4}$. Conditioning on the event we draw a $(\varepsilon, \frac{1}{4})$-good training sample then the probability that $j \notin \Gamma_p$ is at most ϵ by Lemma B.2. Furthermore, with the training sample fixed the activations of each neuron are mutually independent. Let $X = (\mathbf{x}_i)_{i=1}^n$ denote the draw of the training sample and

$$\chi_{\kappa,\epsilon} = \{X : \mathbb{P}_w((B \geq \kappa T) \cup (T = 0)) \leq \epsilon\}$$

the set of $(\varepsilon, \frac{1}{4})$-good training samples. Let $\epsilon = \exp(-cn)$, where c in what follows is a sufficiently small constant. By the assumptions of the lemma

$$\mathbb{P}\left(X \in X_{1/4,\epsilon}\right) \geq 1 - \frac{\delta}{\epsilon} \geq 1 - 2\exp\left(-\frac{n}{1024} + cn\right) \geq 1 - 2\exp\left(-cn\right).$$

For the first result, using the union bound

$$\begin{aligned}
\mathbb{P}\left(|\Gamma_p| = m\right) &\geq \mathbb{P}\left(|\Gamma_p| = m \mid X \in X_{1/4,\epsilon}\right)\mathbb{P}\left(X \in X_{1/4,\epsilon}\right) \\
&\geq \left(1 - m\exp(-cn)\right)\left(1 - 2\exp\left(-cn\right)\right) \\
&\geq 1 - m\exp(-cn)
\end{aligned}$$

as claimed. For the second result observe

$$\begin{aligned}
\mathbb{P}\left(|\Gamma_p| < (1-\alpha)m \mid X \in X_{1/4,\epsilon}\right) &= \mathbb{P}\left(\exists \mathcal{J} \subset [2m]_p, |\mathcal{J}| = \alpha m : j \notin \Gamma_p \; \forall j \in \mathcal{J} \mid X \in X_{\kappa,\epsilon}\right) \\
&\leq \binom{m}{\alpha m}\epsilon^{\alpha m} \\
&\leq \left(\frac{\epsilon e}{\alpha}\right)^{\alpha m}.
\end{aligned}$$

As α is a constant again there is a sufficiently small constant c such that

$$\frac{\epsilon e}{\alpha} = \exp(-cn + 1 + \log(1/\alpha)) \leq \exp(-cn).$$

Therefore, there exists a sufficiently small constant c such that

$$\begin{aligned}
\mathbb{P}\left(|\Gamma_p| \geq (1-\alpha)m\right) &\geq \mathbb{P}\left(|\Gamma_p| \geq (1-\alpha)m \mid X \in X_{1/4,\epsilon}\right)\mathbb{P}\left(X \in X_{1/4,\epsilon}\right) \\
&\geq \left(1 - \exp\left(-cmn\alpha\right)\right)\left(1 - 2\exp\left(-cn\right)\right) \\
&\geq 1 - \exp(-cn)
\end{aligned}$$

as claimed. □

Lemma B.4. *Assume $\gamma + \rho \leq \frac{1}{1.02(n-k)}$, $n - k \geq 2 \times 10^6$, $|\Gamma_p| > 0.99m$ for $p \in \{-1, 1\}$ and $m \geq C\log(n)$ for a sufficiently large constant C. Then with probability at least $1 - \frac{C}{n}$ for all $i \in \mathcal{S}_F$ there exists a $j \in \Theta_{y_i}$ such that $i \in \mathcal{A}_j^{(0)}$.*

Proof. For convenience in what follows we use G_j and B_j for $G_j(0,1)$ and $B_j(0,1)$ respectively. For a neuron to be in Θ_p it must satisfy the following condition,

$$G_j(\gamma + \rho) - B_j(\gamma - \rho) < 1 - \gamma + \rho,$$

or alternatively

$$G_j < \frac{1 - \gamma + \rho}{\gamma + \rho} = \frac{1}{\gamma + \rho} - \frac{\gamma - \rho}{\gamma + \rho}.$$

By our assumptions this is in turn implied by the following condition $G_j < 1.02(n-k) - 1$ or $G_j < 1.01(n-k)$ for $n - k > 100$.

Let $P(i, l)$ be the probability that an arbitrary random neuron is active on points i and l. Consider independently drawing two points from the distribution over points with either positive or negative signal sign component, and let p be the probability that an arbitrary random neuron is active on both points. Similarly let q be the probability that an arbitrary neuron is active on two points which are drawn with opposite signal signs. By rotational invariance of the weight distribution the probability a random neuron activates on a point is $1/2$, therefore $\mathbb{E}(G_j) = n - k$ and furthermore $P(i, i) = 1/2$. In addition, by writing G_j as a sum of indicator functions, expanding, and using the linearity of expectation we have

$$\mathbb{E}(G_j^2) = \sum_{(i,l)\in[n-k]\times[n-k]} P(i, l) = \frac{1}{2}2(n-k) + 2(n-k)^2 q + 2(n-k)(n-k-1)p.$$

Recall i and l index over the clean points. Observe by construction that for $i \not\sim l$ then $-\mathbf{x}_l \overset{d}{=} \mathbf{x}_i$. As a result, using an abuse of notation where the index $-j$ indicates the point $-\mathbf{x}_j$, then for $i \sim l \not\sim j$ we have $P(i, j) = P(i, -l)$. If a neuron activates on $\mathbf{x}_l$ iff it does not activate on $-\mathbf{x}_l$, therefore $P(i, l) + P(i, -l) = P(i, i) = 1/2$ and hence we conclude $p + q = 1/2$. As a result

$$\begin{aligned}
\mathbb{E}(G_j^2) &= (n-k) + 2(n-k)\left((n-k)q + (n-k)p - p\right) \\
&= (n-k) + 2(n-k)\left(\frac{1}{2}(n-k) - p\right) \\
&\leq (n-k) + (n-k)^2.
\end{aligned}$$

As $\mathbb{E}(G_j) = n - k$, it follows that G_j has variance $n - k$. Therefore by Chebyshev's inequality

$$\mathbb{P}(G_j \geq 1.01(n-k)) \leq \frac{10^4}{n-k}.$$

Therefore a given random neuron j satisfies the condition $G_j < 1.01(n-k)$ with failure probability at most $\frac{10^4}{n-k}$. Applying Markov's inequality

$$\mathbb{P}\left(\sum_{j=1}^{2m} \mathbb{1}\left(G_j \geq 1.01(n-k)\right) \geq \frac{2m}{100}\right) \leq \frac{10^6}{2(n-k)}$$

and therefore

$$\mathbb{P}\left(\sum_{j=1}^{2m} \mathbb{1}\left(G_j < 1.01(n-k)\right) \geq 1.998m\right) \geq 1 - \frac{10^6}{2(n-k)}.$$

Now by a Chernoff bound, there exists a small constant $c > 0$ such that with probability at least $1-\exp(-cm)$, a fixed training point is activated by at least $1/3$ of the neurons of each sign. Therefore, using the union bound every training point is activated by at least $m/3$ of the neurons with probability at least $1 - n\exp(-cm) \geq 1 - \exp(-cm)$ using that $\log n \leq O(m)$.

Let $\Lambda := \{j \in [2m] : G_j < 1.01(n-k)\}$ and observe that if $j \in \Gamma_p$ and $j \in \Lambda$ then $j \in \Theta_p$. Condition on two further events in addition to $|\Gamma_p| \geq 0.99m$ for $p \in \{-1, 1\}$: $|\Lambda| \geq 1.998m$ and for all $i \in \mathcal{S}_F$ then $|\mathcal{I}_i(0) \cap \{j : (-1)^j = y_i\}| \geq m/3$ and for any $p \in \{-1, 1\}$. Then with probability one we have for all $i \in \mathcal{S}_F$ that there exists a $j \in \Theta_{y_i}$ such that $i \in \mathcal{A}_j^{(0)}$. The probability that these two conditions hold is at least

$$1 - \frac{C}{n} - \exp(-cm).$$

Supposing that $m \geq C\log(n)$ for a sufficiently large constant C then this probability can in turn be lower bounded as

$$1 - \frac{C}{n}$$

as claimed.

□

The final lemma we provide here states, under mild conditions on the network width, that with high probability every point in the training sample activates a neuron whose output weight matches its label in sign. We will use this to prove our result on non-benign overfitting, which is discussed in Section E.

Lemma B.5. *Let $\delta \in (0,1)$, then if $m \geq \log_2(\frac{2n}{\delta})$ the probability that for all $i \in [2n]$ there exists a $j \in [2m]$ such that $(-1)^j = y_i$ and $i \in \mathcal{A}_j^{(0)}$ is at least $1-\delta$.*

Proof. Observe by the rotational symmetry of the weight distribution that

$$\mathbb{P}(\langle \mathbf{w}_j, \mathbf{x}_i \rangle > 0) = \mathbb{P}(\langle \mathbf{w}_j, \mathbf{e}_1 \rangle > 0) = 1/2.$$

By construction, for each element in the training sample $(\mathbf{x}_i, y_i)$ there are m neurons whose output weight has the same sign as y_i. As the preactivations of $\mathbf{x}_i$ with each neuron are mutually independent from one another, then using the union bound it follows that

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{i=1}^{2n}\{\exists j \in [2m] : (-1)^j = y_i,\ i \in \mathcal{A}_j^{(0)}\}\right) &= 1 - \mathbb{P}\left(\bigcup_{i=1}^{2n}\{\nexists j \in [2m] : (-1)^j = y_i,\ i \in \mathcal{A}_j^{(0)}\}\right) \\
&\geq 1 - 2n\mathbb{P}(\langle \mathbf{w}_j, \mathbf{x}_i \rangle > 0) \\
&= 1 - 2n2^{-m}.
\end{aligned}$$

Setting $\delta \geq 2n2^{-m}$ and rearranging we arrive at the stated result. □

C Supporting Lemmas

C.1 Bounds on activations and preactivations

For any pair of iterations t, t_0 satisfying $t > t_0$, unrolling the GD update rule (4) gives

$$\mathbf{w}_j^{(t_1)} = \mathbf{w}_j^{(t_0)} + (-1)^j \eta \sum_{\ell=1}^{2n} T_{\ell j}(t_0, t) y_\ell \mathbf{x}_\ell.$$

Using (1) and the fact that $\mathbf{n}_i \perp \mathbf{v}$, then for any $i \in [2n]$

$$\begin{aligned}
\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle &= \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle + (-1)^j \eta \sum_{\ell=1}^{2n} T_{\ell j}(t_0, t) y_\ell \langle \mathbf{x}_\ell, \mathbf{x}_i \rangle \\
&= \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle + (-1)^{j+i} \eta \sum_{\ell=1}^{2n} T_{\ell j}(t_0, t)(-1)^{\ell+i} \beta(\ell) \langle \mathbf{x}_\ell, \mathbf{x}_i \rangle \qquad (5) \\
&= \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle + (-1)^{j+i} \beta(i) \eta \sum_{\ell=1}^{2n} T_{\ell j}(t_0, t) \lambda_{i\ell},
\end{aligned}$$

where we define $\lambda_{i\ell} := (-1)^{\ell+i}\beta(i)\beta(\ell)\langle \mathbf{x}_\ell, \mathbf{x}_i \rangle$. Towards the goal of bounding the activation of a neuron with a data point we provide the following results.

Lemma C.1. *Assume $|\langle \mathbf{n}_i, \mathbf{n}_\ell \rangle| \leq \frac{\rho}{1-\gamma}$ for all $i, \ell \in [2n]$ such that $i \neq \ell$.*

1. *If $i = \ell$ then $\lambda_{i\ell} = 1$.*
2. *If $i \neq \ell$, $i \in \mathcal{S}_T$, and $\ell \in \mathcal{S}_F$, then $-(\gamma+\rho) \leq \lambda_{i\ell} \leq -(\gamma-\rho)$.*
3. *If $i \neq \ell$, $i \in \mathcal{S}_F$, and $\ell \in \mathcal{S}_T$, then $-(\gamma+\rho) \leq \lambda_{i\ell} \leq -(\gamma-\rho)$.*
4. *If $i \neq \ell$ and $i, \ell \in \mathcal{S}_T$, then $\gamma - \rho \leq \lambda_{i\ell} \leq \gamma + \rho$.*
5. *If $i \neq \ell$ and $i, \ell \in \mathcal{S}_F$, then $\gamma - \rho \leq \lambda_{i\ell} \leq \gamma + \rho$.*

Proof. Observe by the data model, described in Section B.2, that

$$\langle \mathbf{x}_i, \mathbf{x}_\ell \rangle = (-1)^{\ell+i} \left(\gamma + (1-\gamma)\beta(i)\beta(\ell)\langle \mathbf{n}_i, \mathbf{n}_\ell \rangle\right).$$

Therefore

$$\lambda_{i\ell} = \beta(i)\beta(\ell)\gamma + (1-\gamma)\langle \mathbf{n}_i, \mathbf{n}_\ell \rangle$$

from which the results claimed follow. □

Lemma C.2. *Assume $|\langle \mathbf{x}_i, \mathbf{x}_\ell \rangle| \leq \frac{\rho}{1-\gamma}$ for all $i, \ell \in [2n]$ such that $i \neq \ell$. Then for any $j \in [2m]$ the following are true.*

1. *If $i \in \mathcal{S}_T$, $i \sim j$ then*

$$\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle \geq \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle + \eta \left(T_{ij}(t_0,t) + G_j^{(i)}(t_0,t)(\gamma-\rho) - B_j^{(i)}(t_0,t)(\gamma+\rho) \right)$$
$$\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle \leq \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle + \eta \left(T_{ij}(t_0,t) + G_j^{(i)}(t_0,t)(\gamma+\rho) - B_j^{(i)}(t_0,t)(\gamma-\rho) \right).$$

2. *If $i \in \mathcal{S}_T$, $i \not\sim j$ then*

$$\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle \geq \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle - \eta \left(T_{ij}(t_0,t) + G_j^{(i)}(t_0,t)(\gamma+\rho) - B_j^{(i)}(t_0,t)(\gamma-\rho) \right)$$
$$\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle \leq \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle - \eta \left(T_{ij}(t_0,t) + G_j^{(i)}(t_0,t)(\gamma-\rho) - B_j^{(i)}(t_0,t)(\gamma+\rho) \right).$$

3. *If $i \in \mathcal{S}_F$, $i \sim j$ then*

$$\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle \geq \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle - \eta \left(T_{ij}(t_0,t) - G_j^{(i)}(t_0,t)(\gamma-\rho) + B_j^{(i)}(t_0,t)(\gamma+\rho) \right)$$
$$\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle \leq \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle - \eta \left(T_{ij}(t_0,t) - G_j^{(i)}(t_0,t)(\gamma+\rho) + B_j^{(i)}(t_0,t)(\gamma-\rho) \right).$$

4. *If $i \in \mathcal{S}_F$, $i \not\sim j$ then*

$$\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle \geq \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle + \eta \left(T_{ij}(t_0,t) - G_j^{(i)}(t_0,t)(\gamma+\rho) + B_j^{(i)}(t_0,t)(\gamma-\rho) \right)$$
$$\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle \leq \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle + \eta \left(T_{ij}(t_0,t) - G_j^{(i)}(t_0,t)(\gamma-\rho) + B_j^{(i)}(t_0,t)(\gamma+\rho) \right).$$

Proof. Considering (5) we can further separate the summation term as follows,

$$\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle = \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle + (-1)^{j+i}\beta(i)\eta \left(T_{ij}(t_0,t) + \sum_{\substack{\ell \in \mathcal{S}_T \\ \ell \neq i}} T_{\ell j}(t_0,t)\lambda_{il} + \sum_{\substack{\ell \in \mathcal{S}_F \\ \ell \neq i}} T_{\ell j}(t_0,t)\lambda_{i\ell} \right).$$

Note, with $i \in \mathcal{S}_T$ and $i \sim j$, or $i \in \mathcal{S}_F$ and $i \not\sim j$, then $(-1)^{j+i}\beta(i) = 1$. On the other hand, with $i \in \mathcal{S}_T$ and $i \not\sim j$, or $i \in \mathcal{S}_F$ and $i \sim j$, then $(-1)^{j+i}\beta(i) = -1$. Substituting the relevant bounds on $\lambda_{i\ell}$ provided in Lemma C.1, and observing by definition that $G_j^{(i)}(t_0,t) = \sum_{\ell \in \mathcal{S}_T, \ell \neq i} T_{\ell j}(t_0,t)$ and $B_j^{(i)}(t_0,t) = \sum_{\ell \in \mathcal{S}_T, \ell \neq i} T_{\ell j}(t_0,t)$, one arrives at the results claimed. □

We will often make use of the following similar but more pessimistic bounds on the activations. Recall that ϕ is the ReLU function: $\phi(a) = \max\{a, 0\}$.

Lemma C.3. *For any $j \in [2m]$ and iterations t_0, t with $t_0 \leq t$ the following hold:*

1. *If $i \in \mathcal{S}_T, i \sim j$ then*

$$\phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \geq \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) + \eta T_{ij}(t_0,t) - \eta(\gamma+\rho)B_j(t_0,t) - \eta\phi(\rho-\gamma)G_j^{(i)}(t_0,t)$$
$$\phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \leq \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) + \eta T_{ij}(t_0,t) + \eta(\gamma+\rho)G_j^{(i)}(t_0,t) + \eta\phi(\rho-\gamma)B_j(t_0,t).$$

2. *If $i \in \mathcal{S}_T, i \not\sim j$ then*

$$\phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \geq \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) - \eta T_{ij}(t_0,t) - \eta(\gamma+\rho)G_j^{(i)}(t_0,t) - \eta\phi(\rho-\gamma)B_j(t_0,t)$$
$$\phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \leq \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) - \eta T_{ij}(t_0,t) + \eta(\gamma+\rho)B_j(t_0,t) + \eta\phi(\rho-\gamma)G_j^{(i)}(t_0,t) + \eta.$$

3. If $i \in \mathcal{S}_F, i \sim j$ then

$$\phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \geq \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) - \eta T_{ij}(t_0, t) - \eta(\gamma + \rho) B_j^{(i)}(t_0, t) - \eta\phi(\rho - \gamma) G_j(t_0, t)$$
$$\phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \leq \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) - \eta T_{ij}(t_0, t) + \eta(\gamma + \rho) G_j(t_0, t) + \eta\phi(\rho - \gamma) B_j^{(i)}(t_0, t) + \eta.$$

4. If $i \in \mathcal{S}_F, i \not\sim j$ then

$$\phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \geq \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) + \eta T_{ij}(t_0, t) - \eta(\gamma + \rho) G_j(t_0, t) - \eta\phi(\rho - \gamma) B_j^{(i)}(t_0, t)$$
$$\phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \leq \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) + \eta T_{ij}(t_0, t) + \eta(\gamma + \rho) B_j^{(i)}(t_0, t) + \eta\phi(\rho - \gamma) G_j(t_0, t).$$

The η term in the upper bound for cases 2 and 3 is only necessary if $T_{ij}(t_0, t) > 0$.

We remark that we will often use this result in a setting where $\rho \leq \gamma$. In these cases, the terms that involve $\phi(\rho - \gamma)$ are zero and will be dropped.

Proof. For each of these results, we make use of Lemma C.2, $a \leq \phi(a)$ for all $a \in \mathbb{R}$, and

$$0 \leq G_j^{(i)}(t_0, t_1) \leq G_j(t_0, t_1)$$
$$0 \leq B_j^{(i)}(t_0, t_1) \leq B_j(t_0, t_1)$$

for all i, j, t_0, t_1. We will only prove the inequalities for $i \in \mathcal{S}_T$ here, as the inequalities for $i \in \mathcal{S}_F$ are analogous: statement 4 is just statement 1 with the roles of $G_j(t_0, t)$ and $B_j(t_0, t)$ switched, while statement 3 is the same for statement 2.

For the first inequality in statement 1 we claim it suffices to show

$$\phi(\langle \mathbf{w}_j^{(\tau+1)}, \mathbf{x}_i \rangle) \geq \phi(\langle \mathbf{w}_j^{(\tau)}, \mathbf{x}_i \rangle) + \eta T_{ij}(\tau, \tau+1) - \eta(\gamma+\rho) B_j(\tau, \tau+1) - \eta\phi(\rho-\gamma) G_j^{(i)}(\tau, \tau+1) \tag{6}$$

Indeed, if (6) is true then the result desired follows as

$$\begin{aligned}
\phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) - \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) &= \sum_{\tau=t_0}^{t-1} \phi(\langle \mathbf{w}_j^{(\tau+1)}, \mathbf{x}_i \rangle) - \phi(\langle \mathbf{w}_j^{(\tau)}, \mathbf{x}_i \rangle) \\
&\geq \sum_{\tau=t_0}^{t-1} \Big(\eta T_{ij}(\tau, \tau+1) - \eta(\gamma+\rho) B_j(\tau, \tau+1) \\
&\qquad - \eta\phi(\rho-\gamma) G_j^{(i)}(\tau, \tau+1) \Big) \\
&= \eta T_{ij}(t_0, t) - \eta(\gamma+\rho) B_j(t_0, t) - \eta\phi(\rho-\gamma) G_j^{(i)}(t_0, t).
\end{aligned}$$

In order to prove (6) we bound

$$\begin{aligned}
\phi(\langle \mathbf{w}_j^{(\tau+1)}, \mathbf{x}_i \rangle) &\geq \langle \mathbf{w}_j^{(\tau+1)}, \mathbf{x}_i \rangle \\
&\geq \langle \mathbf{w}_j^{(\tau)}, \mathbf{x}_i \rangle + \eta T_{ij}(\tau, \tau+1) - \eta(\gamma+\rho) B_j(\tau, \tau+1) \\
&\quad - \phi(\rho-\gamma) G_j^{(i)}(\tau, \tau+1).
\end{aligned}$$

This follows from statement 1 in Lemma C.2. From here, we consider two cases: first, if $\langle \mathbf{w}_j^{(\tau)}, \mathbf{x}_i \rangle \geq 0$ then $\langle \mathbf{w}_j^{(\tau)}, \mathbf{x}_i \rangle = \phi(\langle \mathbf{w}_j^{(\tau)}, \mathbf{x}_i \rangle)$ and so (6) clearly holds. Alternatively, if $\langle \mathbf{w}_j^{(\tau)}, \mathbf{x}_i \rangle < 0$ then $T_{ij}(\tau, \tau+1) = 0$, $\phi(\langle \mathbf{w}_j^{(\tau)}, \mathbf{x}_i \rangle) = 0$ and as a result the right-hand-side of (6) is non-positive while the left is non-negative. As such (6) holds trivially.

For the second equality in statement 1 we bound

$$\begin{aligned}
\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle &\leq \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle + \eta \Big(T_{ij}(t_0, t) + G_j^{(i)}(t_0, t)(\gamma+\rho) - B_j^{(i)}(t_0, t)(\gamma-\rho) \Big) \\
&\leq \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) + \eta T_{ij}(t_0, t) + \eta(\gamma+\rho) G_j^{(i)}(t_0, t) + \eta\phi(\rho-\gamma) B_j(t_0, t).
\end{aligned}$$

Since the right-hand side is non-negative, this inequality is true even if we replace the left-hand side by $\phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle)$.

We now proceed to statement 2. For the first inequality, notice that if $\phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) = 0$ then the right-hand side is non-positive and therefore the inequality trivially holds. Otherwise, it must be the case that $\phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) = \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle$. Using statement 2 from Lemma C.2, we obtain the bound

$$\begin{aligned}
\phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) &\geq \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle \\
&\geq \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle - \eta \left(T_{ij}(t_0, t) + G_j^{(i)}(t_0, t)(\gamma + \rho) - B_j^{(i)}(t_0, t)(\gamma - \rho) \right) \\
&\geq \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) - \eta T_{ij}(t_0, t) - \eta(\gamma + \rho) G_j^{(i)}(t_0, t) - \eta\phi(\rho - \gamma) B_j(t_0, t).
\end{aligned}$$

We now turn to the second inequality in statement 2. The corresponding statement from Lemma C.2 yields

$$\begin{aligned}
\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle &\leq \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle - \eta T_{ij}(t_0, t) + \eta(\gamma + \rho) B_j(t_0, t) + \eta\phi(\rho - \gamma) G_j^{(i)}(t_0, t) \\
&\leq \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) - \eta T_{ij}(t_0, t) + \eta(\gamma + \rho) B_j(t_0, t) + \eta\phi(\rho - \gamma) G_j^{(i)}(t_0, t) \qquad (7) \\
&\leq \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) - \eta T_{ij}(t_0, t) + \eta(\gamma + \rho) B_j(t_0, t) + \eta\phi(\rho - \gamma) G_j^{(i)}(t_0, t) + \eta,
\end{aligned}$$

we remark that the reason for the addition of η to the right-hand-side will soon become apparent. The desired inequality holds as long as the right-hand-side is non-negative, we therefore proceed by induction to prove

$$\phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) - \eta T_{ij}(t_0, \tau) + \eta(\gamma + \rho) B_j(t_0, \tau) + \eta\phi(\rho - \gamma) G_j^{(i)}(t_0, t) + \eta \geq 0$$

for $\tau \geq t_0$. The base case $\tau = t_0$ is trivial, assume then that the induction hypothesis holds for some $\tau \geq t_0$. For iteration $\tau + 1$ there are two cases to consider: first, if $\langle \mathbf{w}_j^{(\tau)}, \mathbf{x}_i \rangle < 0$ then $T_{ij}(t_0, \tau+1) = T_{ij}(t_0, \tau)$. In addition, as $B_j(t_0, \tau) \leq B_j(t_0, \tau+1)$ and $G_j^{(i)}(t_0, \tau) \leq G_j^{(i)}(t_0, \tau+1)$ then

$$\begin{aligned}
0 &\leq \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) - \eta T_{ij}(t_0, \tau) + \eta(\gamma + \rho) B_j(t_0, \tau) + \eta\phi(\rho - \gamma) G_j^{(i)}(t_0, \tau) + \eta \\
&\leq \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) - \eta T_{ij}(t_0, \tau + 1) + \eta(\gamma + \rho) B_j(t_0, \tau + 1) + \eta\phi(\rho - \gamma) G_j^{(i)}(t_0, \tau + 1) + \eta
\end{aligned}$$

by the induction hypothesis. Alternatively, if instead $\langle \mathbf{w}_j^{(\tau)}, \mathbf{x}_i \rangle \geq 0$ one may use the second inequality from (7) to conclude that

$$0 \leq \langle \mathbf{w}_j^{(\tau)}, \mathbf{x}_i \rangle \leq \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) - \eta T_{ij}(t_0, \tau) + \eta(\gamma + \rho) B_j(t_0, \tau) + \eta\phi(\rho - \gamma) G_j^{(i)}(t_0, \tau).$$

In addition, as $T_{ij}(t_0, \tau + 1) \leq T_{ij}(t_0, \tau) + 1$ it follows that

$$\begin{aligned}
0 &\leq \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) - \eta T_{ij}(t_0, \tau) + \eta(\gamma + \rho) B_j(t_0, \tau) + \eta\phi(\rho - \gamma) G_j^{(i)}(t_0, \tau) \\
&\leq \phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) - \eta T_{ij}(t_0, \tau + 1) + \eta(\gamma + \rho) B_j(t_0, \tau + 1) + \eta\phi(\rho - \gamma) G_j^{(i)}(t_0, \tau + 1) + \eta
\end{aligned}$$

which completes the induction.

Lastly, we consider the final remark in the statement of the lemma. If $T_{ij}(t_0, t) = 0$ then the right hand side of the second line in (7) is non-negative trivially, so we do not need the additional η term. □

C.2 Useful convergence lemmas

Lemma C.4. *For any iteartions t, t_0 satisfying $t \geq t_0$,*

1. *if $i \in \mathcal{S}_T$ then*

$$y_i f(t, \mathbf{x}_i) \geq y_i f(t_0, \mathbf{x}_i) + \eta(T_i(t_0, t) - (\gamma + \rho) B(t_0, t) - \phi(\rho - \gamma) G^{(i)}(t_0, t) - m),$$

2. *if $i \in \mathcal{S}_F$ then*

$$y_i f(t, \mathbf{x}_i) \geq y_i f(t_0, \mathbf{x}_i) + \eta(T_i(t_0, t) - (\gamma + \rho) G(t_0, t) - \phi(\rho - \gamma) B^{(i)}(t_0, t) - m).$$

Proof. Both statements follow from the bounds provided in Lemma C.3. For Statement 1

$$\begin{aligned}
y_i f(t, \mathbf{x}_i) &= \sum_{j \sim i} \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) - \sum_{j \not\sim i} \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \\
&\geq \sum_{j \sim i} \Big(\phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) + \eta T_{ij}(t_0, t) - \eta(\gamma + \rho) B_j(t_0, t) - \eta \phi(\rho - \gamma) G_j^{(i)}(t_0, t) \Big) \\
&\quad - \sum_{j \not\sim i} \Big(\phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) - \eta T_{ij}(t_0, t) + \eta(\gamma + \rho) B_j(t_0, t) + \eta \phi(\rho - \gamma) G_j^{(i)}(t_0, t) + \eta \Big) \\
&= y_i f(t_0, \mathbf{x}_i) + \eta (T_i(t_0, t) - (\gamma + \rho) B(t_0, t) - \phi(\rho - \gamma) G^{(i)}(t_0, t) - m).
\end{aligned}$$

For Statement 2

$$\begin{aligned}
y_i f(t, \mathbf{x}_i) &= \sum_{j \not\sim i} \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) - \sum_{j \sim i} \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \\
&\geq \sum_{j \not\sim i} \Big(\phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) + \eta T_{ij}(t_0, t) - \eta(\gamma + \rho) G_j(t_0, t) - \eta \phi(\rho - \gamma) B_j^{(i)}(t_0, t) \Big) \\
&\quad - \sum_{j \not\sim i} \Big(\phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) - \eta T_{ij}(t_0, t) + \eta(\gamma + \rho) G_j(t_0, t) + \eta \phi(\rho - \gamma) B_j^{(i)}(t_0, t) + \eta \Big) \\
&\geq y_i f(t_0, \mathbf{x}_i) + \eta (T_i(t_0, t) - (\gamma + \rho) G(t_0, t) - \phi(\rho - \gamma) B^{(i)}(t_0, t) - m).
\end{aligned}$$

□

Lemma C.5. *Let $t \geq t_0$. For $i \in \mathcal{S}_T$,*

$$T_i(t_0, t) \leq \frac{\ell(t_0, \mathbf{x}_i)}{\eta} + (\gamma + \rho) B(t_0, t) + \phi(\rho - \gamma) G^{(i)}(t_0, t) + 3m.$$

For $i \in \mathcal{S}_F$,

$$T_i(t_0, t) \leq \frac{\ell(t_0, \mathbf{x}_i)}{\eta} + (\gamma + \rho) G(t_0, t) + \phi(\rho - \gamma) B^{(i)}(t_0, t) + 3m.$$

Proof. We will show this for $i \in \mathcal{S}_T$; the $i \in \mathcal{S}_F$ case is similar. We proceed by induction on t. If $t = t_0$ this holds trivially because the left-hand side is zero and the right-hand side is positive. Otherwise, assume the desired inequality holds at iteration t. By Lemma C.4 and our assumption on $\ell(t_0, \mathbf{x}_i)$,

$$\begin{aligned}
y_i f(t, \mathbf{x}_i) &\geq y_i f(t_0, \mathbf{x}_i) + \eta (T_i(t_0, t) - (\gamma + \rho) B(t_0, t) - \phi(\rho - \gamma) G^{(i)}(t_0, t) - m) \\
&\geq (1 - a) + \eta (T_i(t_0, t) - (\gamma + \rho) B(t_0, t) - \phi(\rho - \gamma) G^{(i)}(t_0, t) - m).
\end{aligned}$$

We consider two cases:

1. If $\eta (T_i(t_0, t) - (\gamma + \rho) B(t_0, t) - \phi(\rho - \gamma) G^{(i)}(t_0, t) - m) \geq a$ then we see that $\ell(t, \mathbf{x}_i) = 0$. Therefore,

$$\begin{aligned}
T_i(t_0, t+1) &= T_i(t_0, t) \\
&\leq \frac{a}{\eta} + (\gamma + \rho) B(t_0, t) + 3m \\
&\leq \frac{a}{\eta} + (\gamma + \rho) B(t_0, t+1) + 3m
\end{aligned}$$

2. Otherwise, $T_i(t_0, t) \leq \frac{a}{\eta} + (\gamma + \rho) B(t_0, t) + \phi(\rho - \gamma) G^{(i)}(t_0, t) + m$. Since there are only $2m$ neurons, we bound

$$\begin{aligned}
T_i(t_0, t+1) &= T_i(t_0, t) + 2m \\
&\leq \left(\frac{a}{\eta} + (\gamma + \rho) B(t_0, t) + \phi(\rho - \gamma) G^{(i)}(t_0, t) + m \right) + 2m \\
&\leq \frac{a}{\eta} + (\gamma + \rho) B(t_0, t+1) + \phi(\rho - \gamma) G^{(i)}(t_0, t) + 3m.
\end{aligned}$$

□

D Benign overfitting

Assumption 1. *Let $\delta \in (0,1)$. With $n \geq C/\delta$ and $\eta \leq c$ then for benign overfitting we assume the following conditions on the other data and model hyperparameters.*

1. $k < \frac{n}{100}$,
2. $5\sqrt{2\log(2en^2/\delta)/d} \leq \gamma \leq \frac{4}{5n}$
3. $\lambda_w \leq \frac{\eta\gamma}{4}$,
4. $m \geq C\log(n)$
5. $d \geq 3\rho^{-2}\ln\left(\frac{6n^2}{\delta}\right)$ *where* $\rho \leq \frac{\gamma}{5}$

We remark that under these conditions then for sufficiently large n, ρ as defined above clearly satisfies the inequalities $\rho \leq \min\left\{\frac{n-3k}{n+k}\gamma, \frac{1}{6(n-k)}\right\}$ and $\gamma + \rho < \min\left\{\sqrt{\frac{1}{4(n-k)k}}, \frac{1}{n-k}, \frac{1}{99k}, \frac{1}{100}\right\}$. In addition to the assumptions detailed in Assumption 1, for convenience we assume three further conditions hold.

Assumption 2. *In addition to the assumptions detailed in Assumption 1, assume the following conditions hold.*

1. $|\Gamma_p| > 0.99m$ *for* $p \in \{-1, 1\}$.
2. *For all* $i \in \mathcal{S}_F$ *there is* $j \in \Gamma_{y_i}$ *such that* $i \in \mathcal{A}_j^{(0)}$.
3. *For all* $i, l \in [2n]$, $i \neq l$ $|\langle \mathbf{n}_i, \mathbf{n}_l \rangle| \leq \frac{\rho}{1-\gamma}$.

As shown in the following lemma, these two additional conditions hold with high probability over the randomness of the initialization and training set.

Lemma D.1. *The extra conditions of Assumption 2 hold with probability at least* $1 - \delta$.

Proof. Using Lemma B.3, then for sufficiently large n there exists a constant c such that the probability the first condition does not hold is at most $\exp(-cn)$. Alternatively, setting $\delta \geq 3\exp(-cn)$ and rearranging, as long as $n \geq C\ln\left(\frac{3}{\delta}\right)$ then the probability the first condition does not hold is at most $\delta/3$. Using Lemma B.4 then the probability condition two does not hold is at most C/n for some large constant C, therefore as long as $n \geq C/delta$ then the probability the second condition does not hold is at most $\delta/3$. Using Lemma B.1, observing $\frac{\rho}{1-\gamma} > \rho$ and under the conditions of the lemma that $3\rho^{-2}\ln\left(6n^2\right)) > 6$, then as long as

$$d \geq 3\rho^{-2}\ln\left(\frac{6n^2}{\delta}\right)$$

the probability that the third condition does not hold is also at most $\delta/3$. Using the union bound we therefore conclude that all three properties hold with probability at least δ. □

D.1 Proof of Lemma 3.2

Lemma D.2 (Lemma 3.2)**.** *Assume Assumption 2 holds. Suppose further that at some epoch t_0 the loss of every clean point is bounded above by $a \in \mathbb{R}_{\geq 0}$, while the loss of every corrupted point is bounded above by $b \in \mathbb{R}_{\geq 0}$. Then the total number of updates which occurs after this epoch is upper bounded as follows,*

$$G(t_0, t) \leq \frac{2(n-k)}{1 - 4k(n-k)(\gamma+\rho)^2}\left(\frac{a}{\eta} + 3m + 2k(\gamma+\rho)\left(\frac{b}{\eta} + 3m\right)\right)$$
$$B(t_0, t) \leq \frac{2k}{1 - 4k(n-k)(\gamma+\rho)^2}\left(\frac{b}{\eta} + 3m + 2(n-k)(\gamma+\rho)\left(\frac{a}{\eta} + 3m\right)\right)$$

for all $t \geq t_0$.

Proof. From Lemma C.5, $\rho \leq \gamma$, and the assumption on a and b,

$$G(t_0,t) = \sum_{i\in\mathcal{S}_T} T_i(t_0,t) \leq 2(n-k)\left(\frac{a}{\eta} + (\gamma+\rho)B(t_0,t) + 3m\right),$$
$$B(t_0,t) = \sum_{i\in\mathcal{S}_F} T_i(t_0,t) \leq 2k\left(\frac{b}{\eta} + (\gamma+\rho)G(t_0,t) + 3m\right).$$

Substituting these bounds into each other, and as $\gamma+\rho < (4(n-k)k)^{-1/2}$ (Assumption 2), we arrive at the epoch independent bound on the number of updates as claimed in the statement of the theorem. □

D.2 Early training and proof of Lemma 3.3

Lemma D.3. *Assuming Assumption 2, then for $i \in \mathcal{S}_T$ and $j \in \Gamma$ it follows that $\langle \mathbf{w}_j^{(1)}, \mathbf{x}_i\rangle > 0$ iff $i \sim j$. For $i \in \mathcal{S}_F$, $j \in \Theta_p$, and $i \not\sim j$ it follows that $\langle \mathbf{w}_j^{(1)}, \mathbf{x}_i\rangle > 0$ if $i \in \mathcal{A}_j^{(0)}$.*

Proof. Suppose $j \in \Gamma$, $i \sim j$, $i \in \mathcal{S}_T$. Recall from definition of Γ_p that $G_j^{(i)}(0,1)(\gamma-\rho) - B_j^{(i)}(0,1)(\gamma+\rho) \geq \frac{2\lambda_w}{\eta}$. Using Lemma C.2

$$\begin{aligned}
\langle \mathbf{w}_j^{(1)}, \mathbf{x}_i\rangle &\geq \langle \mathbf{w}_j^{(0)}, \mathbf{x}_i\rangle + \eta\left(T_{ij}(0,1) + G_j^{(i)}(0,1)(\gamma-\rho) - B_j^{(i)}(0,1)(\gamma+\rho)\right)\\
&> \langle \mathbf{w}_j^{(0)}, \mathbf{x}_i\rangle + \eta\left(G_j(0,1)(\gamma-\rho) - B_j(0,1)(\gamma+\rho)\right)\\
&\geq \langle \mathbf{w}_j^{(0)}, \mathbf{x}_i\rangle + 2\lambda_w\\
&> \lambda_w.
\end{aligned}$$

On the other hand, if $i \not\sim j$ then again from Lemma C.2

$$\begin{aligned}
\langle \mathbf{w}_j^{(1)}, \mathbf{x}_i\rangle &\leq \langle \mathbf{w}_j^{(0)}, \mathbf{x}_i\rangle - \eta\left(T_{ij}(0,1) + G_j^{(i)}(0,1)(\gamma-\rho) - B_j^{(i)}(0,1)(\gamma+\rho)\right)\\
&\leq \langle \mathbf{w}_j^{(0)}, \mathbf{x}_i\rangle - \eta\left(G_j(0,1)(\gamma-\rho) - B_j(0,1)(\gamma+\rho)\right)\\
&\leq -\lambda_w.
\end{aligned}$$

Now consider $i \in \mathcal{S}_F$, $i \not\sim j$, and $i \in \mathcal{A}_j^{(0)}$. By Lemma C.2 and the definition of Θ,

$$\begin{aligned}
\langle \mathbf{w}_j^{(1)}, \mathbf{x}_i\rangle &\geq \langle \mathbf{w}_j^{(0)}, \mathbf{x}_i\rangle + \eta\left(T_{ij}(0,1) + B_j^{(i)}(0,1)(\gamma-\rho) - G_j^{(i)}(0,1)(\gamma+\rho)\right)\\
&> \eta(1-\gamma+\rho) + \eta\left(B_j(0,1)(\gamma-\rho) - G_j(0,1)(\gamma+\rho)\right)\\
&\geq 0.
\end{aligned}$$

□

Lemma D.4 (Lemma 3.3)**.** *Suppose Assumption 2 holds. Let $j \in \Gamma_p$. Let $0 < t < \mathcal{T}_0$. A point $i \in \mathcal{A}_j^{(t)}$ if one of the following conditions hold:*

1. *$i \in \mathcal{S}_T$ and $i \sim j$*
2. *$i \in \mathcal{S}_F$, $i \not\sim j$, and $i \in \mathcal{A}_j^{(1)}$.*

Furthermore, if one of the following conditions hold, then $i \notin \mathcal{A}_j^{(t)}$:

1. *$i \in \mathcal{S}_T$ and $i \not\sim j$*
2. *$i \in \mathcal{S}_F$, $i \not\sim j$, and $i \notin \mathcal{A}_j^{(1)}$.*

Proof. We proceed by induction. For $t = 1$, the $i \in \mathcal{S}_T$ case was shown in Lemma D.3 and the $i \in \mathcal{S}_F$, $i \not\sim j$ case is clear. Now, suppose the lemma holds for iteration t and consider iteration $t+1$.

First let $i \in \mathcal{S}_F$, $i \not\sim j$. If $i \in \mathcal{A}_j^{(1)}$ then

$$\begin{aligned}\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i \rangle &\geq \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle + \eta \left(T_{ij}(t, t+1) - G_j^{(i)}(t, t+1)(\gamma + \rho) + B_j^{(i)}(t, t+1)(\gamma - \rho) \right) \\ &> \eta \left(1 - (n-k)(\gamma + \rho)\right) \\ &\geq 0.\end{aligned}$$

Here the first line is Lemma C.2, the second line comes from the inductive hypothesis, and the third line comes from $(\gamma + \rho) < \frac{1}{n-k}$ (Assumption 2). If $i \notin \mathcal{A}_j^{(1)}$ then

$$\begin{aligned}\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i \rangle &\leq \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle + \eta \left(T_{ij}(t, t+1) - G_j^{(i)}(t, t+1)(\gamma - \rho) + B_j^{(i)}(t, t+1)(\gamma + \rho) \right) \\ &< \eta \left(-(n-k)(\gamma - \rho) + 2k(\gamma + \rho)\right) \\ &= -\eta((n-3k)\gamma - (n+k)\rho) \\ &\leq 0.\end{aligned}$$

Again, the first line is Lemma C.2, the second line uses the inductive hypothesis, and the fourth line uses $\rho \leq \frac{n-3k}{n+k}\gamma$ (Assumption 2).

Now, let $i \in \mathcal{S}_T$. We again use, in order, Lemma C.2, the inductive hypothesis, and $\rho \leq \frac{n-3k}{n+k}\gamma$. If $i \sim j$ then

$$\begin{aligned}\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i \rangle &\geq \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle + \eta \left(T_{ij}(t, t+1) + G_j^{(i)}(t, t+1)(\gamma - \rho) - B_j^{(i)}(t, t+1)(\gamma + \rho) \right) \\ &> \eta(1 + \rho - \gamma) + \eta((n-k)(\gamma - \rho) - 2k(\gamma + \rho)) \\ &> 0.\end{aligned}$$

If $i \not\sim j$ then

$$\begin{aligned}\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i \rangle &\leq \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle - \eta \left(T_{ij}(t, t+1) + G_j^{(i)}(t, t+1)(\gamma - \rho) - B_j^{(i)}(t, t+1)(\gamma + \rho) \right) \\ &< -\eta((n-k)(\gamma - \rho) - 2k(\gamma + \rho)) \\ &= -\eta((n-3k)\gamma - (n+k)\rho) \\ &\leq 0.\end{aligned}$$

□

Lemma D.5. *Suppose Assumption 2 holds. For all $t_0 \leq t_1 < \mathcal{T}_0$,*

$$G_j(t_0, t_1) \leq (n-k)(t_1 - t_0 + 2) + \frac{1}{\gamma - \rho}$$

Proof. First we claim that for all $i \in \mathcal{S}_T$, $j \not\sim i$, and $t < \mathcal{T}_0$,

$$\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle \leq \lambda_w + 2\eta k(\gamma + \rho).$$

We prove the claim by induction. The base case $t = 0$ follows because $\langle \mathbf{w}_j^{(0)}, \mathbf{x}_i \rangle \leq \lambda_w$. Now suppose it is true at iteration t. If $\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle > 0$ then by Lemma C.2,

$$\begin{aligned}\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i \rangle &\leq \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle - \eta T_{ij}(t, t+1) + \eta(\gamma + \rho) B_j(t, t+1) \\ &\leq \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle - \eta(1 - 2k(\gamma + \rho)) \\ &\leq \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle\end{aligned}$$

using $\gamma + \rho < \frac{1}{2k}$. From this, the claim follows. Otherwise,

$$\begin{aligned}\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i \rangle &\leq \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle - \eta T_{ij}(t, t+1) + \eta(\gamma + \rho) B_j(t, t+1) \\ &\leq 2\eta k(\gamma + \rho).\end{aligned}$$

We now turn to the statement of the lemma, again proceeding by induction. The base case $t_1 = t_0$ is clear. Otherwise, we consider two cases:

1. If $G_j(t_0,t_1) > \frac{1+2k(t_1-t_0+1)(\gamma+\rho)}{\gamma-\rho}$ then for all $i \in \mathcal{S}_T$ and $j \not\sim i$, by Lemma C.2,

$$\begin{aligned}\langle \mathbf{w}_j^{(t_1)}, \mathbf{x}_i\rangle &\le \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i\rangle - \eta\left(T_{ij}(t_0,t_1) + G_j^{(i)}(t_0,t_1)(\gamma-\rho) - B_j^{(i)}(t_0,t_1)(\gamma+\rho)\right)\\ &\le \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i\rangle - \eta\left(G_j(t_0,t_1)(\gamma-\rho) - B_j(t_0,t_1)(\gamma+\rho)\right)\\ &< (\lambda_w + 2\eta k(\gamma+\rho)) + 2\eta k(t_1-t_0)(\gamma+\rho) - \eta G_j(t_0,t_1)(\gamma-\rho)\\ &\le 0\end{aligned}$$

by the claim and $\lambda_w < \eta$ (Assumption 2). Therefore, $G_j(t_0,t_1+1) \le G_j(t_0,t_1) + (n-k)$.

2. If $G_j(t_0,t_1) \le \frac{1+2k(t_1-t_0+1)(\gamma+\rho)}{\gamma-\rho}$, then

$$\begin{aligned}G_j(t_0,t_1+1) &\le \frac{1+2k(t_1-t_0+1)(\gamma+\rho)}{\gamma-\rho} + 2(n-k)\\ &\le \frac{1}{\gamma-\rho} + \frac{2k(t_1-t_0+1)(n-k)}{2k} + 2(n-k)\\ &\le \frac{1}{\gamma-\rho} + (t_1-t_0+3)(n-k).\end{aligned}$$

□

Lemma D.6. *Suppose Assumption 2 holds. For all $t < \mathcal{T}_0$, $i \in \mathcal{S}_F$, and $i \sim j$,*

$$\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i\rangle \le (\lambda_w + 2\eta k(\gamma-\rho)) + 2\eta(\gamma+\rho)(n-k) + \frac{\eta(\gamma+\rho)}{\gamma-\rho}$$

Proof. Consider $t < \mathcal{T}_0$. We consider three cases

1. If $t = 0$ then
$$\langle \mathbf{w}_j^{(0)}, \mathbf{x}_i\rangle \le \lambda_w.$$

2. If $\langle \mathbf{w}_j^{(t-1)}, \mathbf{x}_i\rangle \le 0$ then by Lemma C.2
$$\begin{aligned}\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i\rangle &\le \langle \mathbf{w}_j^{(t-1)}, \mathbf{x}_i\rangle\\ &\quad - \eta\left(T_{ij}(t-1,t) - G_j^{(i)}(t-1,t)(\gamma+\rho) + B_j^{(i)}(t-1,t)(\gamma-\rho)\right)\\ &\le 2\eta k(\gamma-\rho).\end{aligned}$$

3. If $\langle \mathbf{w}_j^{(t-1)}, \mathbf{x}_i\rangle > 0$ then let $t' < t$ be the smallest iteration such that $\langle \mathbf{w}_j^{(\tau)}, \mathbf{x}_i\rangle > 0$ for all $t' \le \tau < t$. By Lemma C.2, Lemma D.5, and the previous two cases above,
$$\begin{aligned}\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i\rangle &\le \langle \mathbf{w}_j^{(t')}, \mathbf{x}_i\rangle - \eta\left(T_{ij}(t',t) - G_j^{(i)}(t',t)(\gamma+\rho) + B_j^{(i)}(t',t)(\gamma-\rho)\right)\\ &\le (\lambda_w + 2\eta k(\gamma-\rho))\\ &\quad - \eta\left([1-(\gamma+\rho)(n-k)](t-t') - 2(\gamma+\rho)(n-k) - \frac{\gamma+\rho}{\gamma-\rho}\right).\end{aligned}$$
By $\gamma+\rho < \frac{1}{n-k}$ (Assumption 2) we conclude
$$\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i\rangle \le (\lambda_w + 2\eta k(\gamma-\rho)) + 2\eta(\gamma+\rho)(n-k) + \frac{\eta(\gamma+\rho)}{\gamma-\rho}.$$

□

D.3 Proof of Lemma 3.4

Lemma D.7 (Lemma 3.4)**.** *Suppose Assumption 2 holds. There is an iteration $\mathcal{T}_1 < \mathcal{T}_0$ during training and expressions C_1, C_2, and C_3 where the following hold:*

1. *For all $p \in \{-1,1\}$, $j \in \Gamma_p$, $i \sim j$, and $i \in \mathcal{S}_T$,*
$$\langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i\rangle \ge \frac{3}{4m}.$$

2. *For all* $p \in \{-1, 1\}$, $j \in \Gamma_p$, $i \not\sim j$, *and* $i \in \mathcal{S}_T$,

$$\langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i \rangle \leq -\frac{3n\gamma}{4m}.$$

3. *For all* $i \in \mathcal{S}_T$,

$$\ell(\mathcal{T}_1, \mathbf{x}_i) \leq \frac{1}{3}.$$

Furthermore,

$$\mathcal{T}_1 = \frac{1}{1.03\eta m(1 + (\gamma + \rho)(n - k))} + O(1)$$

Proof. Fix $i \in \mathcal{S}_T$. At every iteration $1 \leq t < \mathcal{T}_0$, we bound

$$\begin{aligned}
\ell(t, \mathbf{x}_i) - \ell(t+1, \mathbf{x}_i) &= y_i[f(t+1, \mathbf{x}_i) - f(t, \mathbf{x}_i)] \\
&= \sum_{j=1}^{2m} (-1)^{i+j}[\phi(\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i \rangle) - \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle)] \\
&= \sum_{j \notin \Gamma_{(-1)^{i+1}}} (-1)^{i+j}[\phi(\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i \rangle) - \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle)] \\
&\leq \eta \sum_{j \notin \Gamma_{(-1)^{i+1}}} T_{ij}(t, t+1) + (\gamma + \rho) G_j^{(i)}(t, t+1) \\
&= \eta \left(\sum_{j \in \Gamma_{(-1)^i}} T_{ij}(t, t+1) + (\gamma + \rho) G_j^{(i)}(t, t+1) \right) \\
&\quad + \eta \left(\sum_{j \notin \Gamma} T_{ij}(t, t+1) + (\gamma + \rho) G_j^{(i)}(t, t+1) \right) \\
&\leq \eta 0.99m[1 + (\gamma + \rho)(n - k)] + 0.02\eta m[1 + 2(\gamma + \rho)(n - k)],
\end{aligned}$$

where we use in order: $t < \mathcal{T}_0$, the definition of $f(t, \mathbf{x})$, Lemma D.4, Lemma C.3, $\Gamma_{-1} \cap \Gamma_1 = \emptyset$, and Lemma D.4 again. We also use $|\Gamma_p| \geq 0.99m$ (Assumption 2). We further simplify this bound to conclude

$$\ell(t+1, \mathbf{x}_i) - \ell(t, \mathbf{x}_i) \leq 1.03\eta m[1 + (\gamma + \rho)(n - k)]$$

Additionally, we bound

$$\begin{aligned}
\ell(1, \mathbf{x}_i) &= 1 - y_i f(1, \mathbf{x}_i) \\
&= 1 - \sum_{j=1}^{2m} (-1)^{i+j} \phi(\langle \mathbf{w}_j^{(1)}, \mathbf{x}_i \rangle) \\
&\geq 1 - \sum_{i \sim j} \phi(\langle \mathbf{w}_j^{(1)}, \mathbf{x}_i \rangle) \\
&\geq 1 - \sum_{i \sim j} [\phi(\langle \mathbf{w}_j^{(0)}, \mathbf{x}_i \rangle) + \eta T_{ij}(0, 1) + \eta(\gamma + \rho) G_j^{(i)}(0, 1)] \\
&\geq 1 - m[\lambda_w + \eta + 2\eta(\gamma + \rho)(n - k)].
\end{aligned}$$

Therefore as long as

$$t < \frac{1 - m[\lambda_w + \eta + 2\eta(\gamma + \rho)(n - k)]}{1.03\eta m[1 + (\gamma + \rho)(n - k)]} + 1 = \frac{1}{1.03\eta m[1 + (\gamma + \rho)(n - k)]} + O(1),$$

then

$$\begin{aligned}
\ell(t, \mathbf{x}_i) &= \ell(1, \mathbf{x}_i) + \sum_{t'=1}^{t} \ell(t'+1, \mathbf{x}_i) - \ell(t', \mathbf{x}_i) \\
&\geq 1 - m[\lambda_w + \eta + 2\eta(\gamma + \rho)(n - k)] - 1.03(t' - 1)\eta m[1 + (\gamma + \rho)(n - k)] \\
&> 0.
\end{aligned}$$

Notice that this does not depend on i or p. Therefore we can let $\mathcal{T}_1$ be the largest integer satisfying this bound for t and bound $\ell(\mathcal{T}_1, \mathbf{x}_\ell) > 0$ for all $l \in \mathcal{S}_T$. To verify that $\mathcal{T}_1 < \mathcal{T}_0$, consider $i \in \mathcal{S}_F$:

$$\begin{aligned}
y_i f(t, \mathbf{x}_i) &= \sum_{j=1}^{2m} -(-1)^{i+j} \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \\
&= \sum_{i \not\sim j} [(\phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) - \phi(\langle \mathbf{w}_j^{(0)}, \mathbf{x}_i \rangle)) + \phi(\langle \mathbf{w}_j^{(0)}, \mathbf{x}_i \rangle)] \\
&\leq \sum_{i \not\sim j} \Big(T_{ij}(0,t) + (\gamma + \rho) B_j^{(i)}(0,t) + \lambda_w \Big) \\
&\leq \eta m t [1 + 2(\gamma + \rho)k] + m\lambda_w
\end{aligned}$$

This is less than 1 for all $t < \mathcal{T}_1$ since $k \leq \frac{n}{3}$ (Assumption 2).

Now, fix $i \in \mathcal{S}_T$ again. For $i \sim j$, we then can use Lemma C.2 and Lemma D.4:

$$\begin{aligned}
\langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i \rangle &\geq \langle \mathbf{w}_j^{(1)}, \mathbf{x}_i \rangle + \eta \Big(T_{ij}(1, \mathcal{T}_1) + (\gamma - \rho) G_j^{(i)}(1, \mathcal{T}_1) - (\gamma + \rho) B_j^{(i)}(1, \mathcal{T}_1) \Big) \\
&\geq 0 + \eta(\mathcal{T}_1 - 2)(1 + (n - k - 1)(\gamma - \rho) - 2k(\gamma + \rho)) \\
&= \frac{1 + (n - k - 1)(\gamma - \rho) - 2k(\gamma + \rho)}{1.03m[1 + (\gamma + \rho)(n - k)]} + O(\eta) \\
&\geq \frac{1}{m} - \frac{2\rho(n - k) - 2k(\gamma + \rho) - (\gamma + \rho) - 2k(\gamma + \rho)}{m(1 + (\gamma + \rho)(n - k))} + O(\eta) \\
&\geq \frac{1}{m} - \frac{1/6 + 4/99 + 1/100}{m} + O(\eta) \\
&\geq \frac{3}{4m}.
\end{aligned}$$

using $\rho \leq \frac{1}{6(n-k)}$, η is sufficiently small, and $\gamma + \rho < \min\left\{\frac{1}{99k}, \frac{1}{100}\right\}$ (Assumption 2). Now assume $i \not\sim j$. Using Lemma C.2 and Lemma D.4 we can bound

$$\begin{aligned}
\langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i \rangle &\leq \langle \mathbf{w}_j^{(1)}, \mathbf{x}_i \rangle - \eta \Big(T_{ij}(1, \mathcal{T}_1) + (\gamma - \rho) G_j^{(i)}(1, \mathcal{T}_1) - (\gamma + \rho) B_j^{(i)}(1, \mathcal{T}_1) \Big) \\
&\leq 0 - \eta(\mathcal{T}_1 - 2)((n - k)(\gamma - \rho) - 2k(\gamma + \rho)) \\
&= -\frac{(n - k)(\gamma - \rho) - 2k(\gamma + \rho)}{1.03m[1 + (\gamma + \rho)(n - k)]} + O(\eta) \\
&\leq -\frac{(n - 3k)\gamma - (n + k)\rho}{1.03m} + O(\eta) \\
&\leq -\frac{0.97n\gamma - \frac{1.01}{5}n}{1.03m} + O(\eta) \\
&\leq -\frac{3n\gamma}{4m}.
\end{aligned}$$

using η is sufficiently small and $k \leq \frac{n}{100}$ and $\rho \leq \frac{\gamma}{5}$ (Assumption 2). Likewise, for $1 \leq t < \mathcal{T}_1$,

$$\begin{aligned}
\ell(t, \mathbf{x}_i) - \ell(t+1, \mathbf{x}_i) &= y_i[f(t+1, \mathbf{x}_i) - f(t, \mathbf{x}_i)] \\
&= \sum_{j=1}^{2m} (-1)^{i+j} [\phi(\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i \rangle) - \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle)] \\
&= \sum_{j \notin \Gamma_{(-1)^{i+1}}} (-1)^{i+j} [\phi(\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i \rangle) - \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle)] \\
&= \eta \left(\sum_{j \in \Gamma_{(-1)^i}} [\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i \rangle - \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle] \right) \\
&\quad + \eta \left(\sum_{j \notin \Gamma} (-1)^{i+j} [\phi(\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i \rangle) - \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle)] \right) \\
&\geq \eta \left(\sum_{j \in \Gamma_{(-1)^i}} T_{ij}(t_0, t) + G_j^{(i)}(t_0, t)(\gamma - \rho) - B_j^{(i)}(t_0, t)(\gamma + \rho) \right) \\
&\quad + \eta \left(\sum_{j \notin \Gamma} T_{ij}(t, t+1) - (\gamma + \rho) B_j(t, t+1) \right) \\
&\geq 0.99\eta m[1 + (\gamma - \rho)(n - k - 1) - 2(\gamma + \rho)k] - 0.04\eta m(\gamma + \rho)k \\
&\geq 0.99\eta m[1 + (\gamma - \rho)(n - k - 1)] - 2.02\eta m(\gamma + \rho)k
\end{aligned}$$

In the first six lines we use: $t < \mathcal{T}_0$, the definition of $f(t, \mathbf{x})$, Lemma D.4, Lemma C.2 and Lemma C.3, Lemma D.4 again, and $|\Gamma_p| \geq 0.99m$ (Assumption 2), respectively.

We also bound

$$\begin{aligned}
\ell(1, \mathbf{x}_i) &= 1 - y_i f(1, \mathbf{x}_i) \\
&= 1 - \sum_{j=1}^{2m} (-1)^{i+j} \phi(\langle \mathbf{w}_j^{(1)}, \mathbf{x}_i \rangle) \\
&\leq 1 + \sum_{i \not\sim j} \phi(\langle \mathbf{w}_j^{(1)}, \mathbf{x}_i \rangle) \\
&\leq 1 + \sum_{i \not\sim j} [\phi(\langle \mathbf{w}_j^{(0)}, \mathbf{x}_i \rangle) - \eta T_{ij}(0, 1) + \eta(\gamma + \rho) B_j(0, 1) + \eta] \\
&\leq 1 + m[\lambda_w + \eta + 2\eta(\gamma + \rho)k].
\end{aligned}$$

Combining these two bounds we see that

$$\begin{aligned}
\ell(\mathcal{T}_1, \mathbf{x}_i) &\leq \ell(1, \mathbf{x}_i) - \eta(\mathcal{T}_1 - 1)(0.99m(1 + (\gamma - \rho)(n - k - 1)) - 2.02m(\gamma + \rho)k) \\
&\leq 1 - \frac{0.99m[1 + (\gamma - \rho)(n - k - 1)] - 2.02m(\gamma + \rho)k}{1.03m(1 + (\gamma + \rho)(n - k))} + O(\eta) \\
&\leq \frac{0.99((\gamma + \rho) + 2\rho(n - k - 1)) + 0.08 + \frac{4.04}{99}}{1.03m(1 + (\gamma + \rho)(n - k))} + O(\eta) \\
&\leq \frac{0.99(\frac{1}{100} + \frac{1}{6} + 0.08 + \frac{4.04}{99})}{1.03} + O(\eta) \\
&\leq \frac{1}{3}.
\end{aligned}$$

at this iteration, as desired. Here we use $(\gamma + \rho) \leq \min\left\{\frac{1}{99k}, \frac{1}{100}, \frac{1}{n-k}\right\}$, η is sufficiently small, and $\rho \leq \frac{1}{6(n-k)}$ (Assumption 2) $\square$

D.4 Late training

Lemma D.8. *Suppose Assumption 2 holds. Fix $\varepsilon > 0$. We will say a neuron is* aligned *(at iteration t) if*

$$(-1)^j \operatorname{sgn}\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i\rangle = y_i$$

for all $i \in \mathcal{S}_T$. For $t \geq \mathcal{T}_1$, if

$$B(\mathcal{T}_1, t) \leq \frac{5\varepsilon n}{8\eta}$$

than at least $(1-\varepsilon)m$ neurons in each Γ_p will be aligned.

Proof. Let $p \in \{-1, 1\}$ and t be such that εm different neurons in Γ_p are unaligned at iteration t. For any neuron index $j \in \Gamma_p$ and $i \in \mathcal{S}_T$, we can use Lemma C.2 to bound

$$\begin{aligned}
(-1)^{i+j}\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i\rangle &\geq (-1)^{i+j}\langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i\rangle + \eta T_{ij}(\mathcal{T}_1, t) \\
&\quad + \eta(\gamma-\rho)G_j^{(i)}(\mathcal{T}_1, t) - \eta(\gamma+\rho)B_j^{(i)}(\mathcal{T}_1, t) \\
&= \min\left\{\frac{3}{4m}, \frac{3n\gamma}{4m}\right\} - \eta(\gamma+\rho)B_j(\mathcal{T}_1, t) \\
&\geq \frac{3n\gamma}{4m} - \eta(\gamma+\rho)B_j(\mathcal{T}_1, t).
\end{aligned}$$

Since $n\gamma < 1$ (Assumption 2), we see that $\min\{\frac{3}{4m}, \frac{3n\gamma}{4m}\} = \frac{3n\gamma}{4m}$. If the lower bound above is positive then $(-1)^j \operatorname{sgn}\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i\rangle = y_i$. Therefore, if a neuron j is unaligned then $B_j(\mathcal{T}_1, t) \geq \frac{3n\gamma}{4\eta m(\gamma+\rho)} \geq \frac{5n}{8\eta m}$ (using $\rho \leq \frac{\gamma}{5}$ from Assumption 2). If there are εm unaligned neurons, then

$$B(\mathcal{T}_1, t) = \sum_{j=1}^{2m} B_j(\mathcal{T}_1, t) \geq \frac{5\varepsilon n}{8\eta}.$$

□

Denote the first iteration after $\mathcal{T}_1$ where more than εm neurons in one of the Γ_p are unaligned as $\mathcal{T}^\varepsilon$. If no such iteration exists, let $\mathcal{T}^\varepsilon = \infty$. We will eventually show that indeed $\mathcal{T}^\varepsilon = \infty$, by showing that the training process reaches zero loss before such an iteration can happen.

Lemma D.9. *Assume Assumption 2 holds and also $\gamma + \rho < \frac{0.99(1-\varepsilon)}{4k}$. There is an iteration $\mathcal{T}_2 \geq \mathcal{T}_1$ so that for all iterations t satisfying $\mathcal{T}_2 \leq t < \mathcal{T}^\varepsilon$ and all $i \in \mathcal{S}_T$,*

$$\ell(t, \mathbf{x}_i) \leq 4\eta(\gamma+\rho)km.$$

Furthermore, we can choose $\mathcal{T}_2$ so that

$$\mathcal{T}_2 - \mathcal{T}_1 \leq \frac{1}{3\eta m(0.99(1-\varepsilon) - 4k(\gamma+\rho))} + 1.$$

Proof. Fix $i \in \mathcal{S}_T$ and $t < \mathcal{T}^\varepsilon - 1$. Suppose $\ell(t, \mathbf{x}_i) > 0$. Using Lemma C.3 and $t < \mathcal{T}^\varepsilon$,

$$\begin{aligned}
y_i f(t+1, \mathbf{x}_i) - y_i f(t, \mathbf{x}_i) &= \sum_{j=1}^{2m} (-1)^{i+j}(\phi(\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i\rangle) - \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i\rangle)) \\
&\geq \sum_{j=1}^{2m} \eta(T_{ij}(t, t+1) - (\gamma+\rho)B_j(t, t+1)) \\
&\geq \eta 0.99(1-\varepsilon)m - 4\eta mk(\gamma+\rho).
\end{aligned}$$

Therefore,

$$\ell(t+1, \mathbf{x}_i) \leq \min\{\ell(t, \mathbf{x}_i) - \eta m(0.99(1-\varepsilon) - 4k), 0\}.$$

By Lemma D.7, the loss of each clean point at $\mathcal{T}_1$ is at most $\frac{1}{3}$, so each clean point reaches zero loss in at most

$$\left\lceil \frac{1}{3\eta m(0.99(1-\varepsilon) - 4k(\gamma+\rho))} \right\rceil$$

iterations.

Now suppose $\ell(t, \mathbf{x}_i) = 0$. We similarly argue

$$\begin{aligned} y_i f(t+1, \mathbf{x}_i) - y_i f(t, \mathbf{x}_i) &= \sum_{j=1}^{2m} (-1)^{i+j} (\phi(\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i \rangle) - \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle)) \\ &\geq \sum_{j=1}^{2m} \eta (T_{ij}(t, t+1) - B_j(t, t+1)) \\ &\geq -4\eta mk. \end{aligned}$$

This implies $\ell(t+1, \mathbf{x}_i) \leq 4\eta mk$. By induction, we see that if $t < \mathcal{T}^\varepsilon$ and $\ell(t, \mathbf{x}_i) \leq 4\eta mk$, then $\ell(t+1, \mathbf{x}_i) \leq 4\eta mk$. □

Lemma D.10. *Assume Assumption 2 holds and $\gamma + \rho < \frac{0.99(1-\varepsilon)}{4k}$. For all t_1, t_2 satisfying $\mathcal{T}_2 \leq t_1 \leq t_2 < \mathcal{T}^\varepsilon$ and $i \in \mathcal{S}_T$,*

$$S_i(t_1, t_2) \leq \frac{4(t_2 - t_1)(\gamma + \rho)k + 4k + 3}{0.99(1 - \varepsilon)}.$$

Proof. Recall Lemma C.5 (restated in this setting, using Lemma D.9):

$$T_i(t_1, t_2) \leq \frac{4\eta mk}{\eta} + (\gamma + \rho) B(t_1, t_2) + 3m.$$

Using $t < \mathcal{T}^\varepsilon$ we bound

$$\begin{aligned} 0.99(1 - \varepsilon) m S_i(t_1, t_2) &\leq T_i(t_1, t_2) \\ &\leq \frac{4\eta mk}{\eta} + (\gamma + \rho) B(t_1, t_2) + 3m \\ &\leq 4mk + 4(t_2 - t_1)(\gamma + \rho) mk + 3m. \end{aligned}$$

From this, the desired inequality follows. □

Lemma D.11. *Assume Assumption 2 holds and $\gamma + \rho \leq \min\left\{\frac{0.99(1-\varepsilon)}{4k}, \sqrt{\frac{0.99(1-\varepsilon)}{8(n-k)k}}\right\}$. Let $i \in \mathcal{S}_F$. Suppose there is $j \not\sim i$ such that*

$$\langle \mathbf{w}_j^{(\mathcal{T}_2)}, \mathbf{x}_i \rangle > \eta \frac{2(n-k)(\gamma + \rho)(4k+3)}{0.99(1-\varepsilon)}.$$

Then for all t satisfying $\mathcal{T}_2 \leq t < \mathcal{T}^\varepsilon$, $\langle \mathbf{w}_{j'}^{(t)}, \mathbf{x}_i \rangle > 0$ for some neuron j' depending on t.

Proof. Let $\tau_0 \geq \mathcal{T}_2$ be the first iteration after $\mathcal{T}_2$ where $\ell(t, \mathbf{x}_i) = 0$. We will show by induction that for all t satisfying $\mathcal{T}_2 \leq t \leq \tau_0$ that $\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle > 0$. The case $t = \mathcal{T}_2$ follows immediately by the assumption of the lemma. Otherwise, assume $\langle \mathbf{w}_j^{(t')}, \mathbf{x}_i \rangle > 0$ for all $\mathcal{T}_2 \leq t' < t$. By Lemma C.2

and Lemma D.10,

$$\begin{aligned}
\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i \rangle &\geq \langle \mathbf{w}_j^{(\mathcal{T}_2)}, \mathbf{x}_i \rangle \\
&\quad + \eta \left(T_{ij}(\mathcal{T}_2, t+1) - G_j^{(i)}(\mathcal{T}_2, t+1)(\gamma+\rho) + B_j^{(i)}(\mathcal{T}_2, t+1)(\gamma-\rho) \right) \\
&\geq \langle \mathbf{w}_j^{(\mathcal{T}_2)}, \mathbf{x}_i \rangle + \eta(t+1-\mathcal{T}_2) - \eta(\gamma+\rho) \sum_{i \in \mathcal{S}_T} S_i(\mathcal{T}_2, t+1) \\
&\geq \langle \mathbf{w}_j^{(\mathcal{T}_2)}, \mathbf{x}_i \rangle + \eta(t+1-\mathcal{T}_2) \\
&\quad - \eta 2(n-k)(\gamma+\rho) \left(\frac{4(t+1-\mathcal{T}_2)(\gamma+\rho)k + 4k + 3}{0.99(1-\varepsilon)} \right) \\
&= \eta(t+1-\mathcal{T}_2) \left(1 - \frac{8(n-k)k(\gamma+\rho)^2}{0.99(1-\varepsilon)} \right) + \langle \mathbf{w}_j^{(\mathcal{T}_2)}, \mathbf{x}_i \rangle \\
&\quad - \eta \frac{2(n-k)(\gamma+\rho)(4k+3)}{0.99(1-\varepsilon)} \\
&> 0.
\end{aligned}$$

We now continue the induction past τ_0. If $\ell(t, \mathbf{x}_i) = 0$, then point i clearly activates some neuron. Let τ_1 be the first iteration after τ_0 where $\ell(t, \mathbf{x}_i) > 0$. By Lemma C.3

$$\begin{aligned}
y_i f(\tau_1, \mathbf{x}_i) - y_i f(\tau_1 - 1, \mathbf{x}_i) &= \sum_{j'=1}^{2m} -(-1)^{i+j} (\phi(\langle \mathbf{w}_j^{(\tau_1)}, \mathbf{x}_i \rangle) - \phi(\langle \mathbf{w}_j^{(\tau_1 - 1)}, \mathbf{x}_i \rangle)) \\
&\geq \sum_{j=1}^{2m} \eta(T_{ij}(t, \tau_1) - G_j(t, \tau_1 - 1)) \\
&\geq -4\eta m(n-k).
\end{aligned}$$

This means

$$\sum_{j' \not\sim i} (-1)^j \phi(\langle \mathbf{w}_{j'}^{(\tau_1)}, \mathbf{x}_i \rangle) \geq y_i f(\tau_1, \mathbf{x}_i) \geq 1 - 4\eta m(n-k)$$

and there is some j' satisfying

$$\phi(\langle \mathbf{w}_{j'}^{(\tau_1)}, \mathbf{x}_i \rangle) \geq \frac{1}{m} - 4\eta(n-k) > \eta \frac{2(n-k)(\gamma+\rho)(4k+3)}{0.99(1-\varepsilon)},$$

assuming η is sufficiently small (Assumption 2). We can run the original induction argument with τ_1 replacing $\mathcal{T}_2$ and $\tau_2 = \min\{t \geq \tau_2 : \ell(t, \mathbf{x}_i) = 0\}$ replacing τ_0 to verify the conclusion for $\tau_1 \leq t \leq \tau_2$. By switching back and forth between these two arguments, we can show that point i activates some neuron for all $\mathcal{T}_2 \leq t < \mathcal{T}^\varepsilon$. □

Lemma D.12. *If Assumption 2 holds, the training process reaches loss.*

Proof. In this proof, let $\varepsilon = \frac{1}{5}$. The conditions of Lemma D.9, Lemma D.10, and Lemma D.11 hold because $\gamma + \rho \leq \min\left\{ \frac{0.99/5}{k}, \sqrt{\frac{0.99/5}{2k(n-k)}} \right\}$.

By Lemma D.2, there is a finite bound on the number of updates, independent of the number of iterations spent training. If we carry out the training procedure for infinitely many iterations, there must be some iteration where we make no updates. Since the training procedure is deterministic, we will not make any updates after this point, and we will have converged. It remains to show that this convergence results in zero training loss. The only way for a point to not update any neurons is for that point's loss to be zero or for that point to activate no neurons.

Lemma D.8 and Lemma D.11 say, under certain conditions, that every clean point and every corrupted point activates some neuron for each iteration $t \leq \mathcal{T}^\varepsilon$. We need only to verify that these conditions hold and that $B(\mathcal{T}_1, t)$ remains below the limitation set in Lemma D.8.

We apply Lemma D.2 starting at $t_0 = \mathcal{T}_1$. By Lemma D.7, $\ell(\mathcal{T}_1, \mathbf{x}_i) \le \frac{1}{3}$ for all $i \in \mathcal{S}_T$. Using Lemma D.6, for $i \in \mathcal{S}_F$,

$$\begin{aligned}\ell(\mathcal{T}_1, \mathbf{x}_i) &\le 1 - y_i \sum_{j=1}^{2m} (-1)^j \phi(\langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i \rangle) \\ &\le 1 + \sum_{j \sim i}^{2m} \phi(\langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i \rangle) \\ &\le 1 + O(\eta).\end{aligned}$$

With these bounds, Lemma D.2 shows that for all $t \ge \mathcal{T}_1$,

$$\begin{aligned}B(\mathcal{T}_1, t) &\le \frac{2k}{1 - 4k(n-k)(\gamma+\rho)^2} \left(\frac{1}{\eta} + 2(n-k)(\gamma+\rho)\frac{1}{3\eta} \right) + O(1) \\ &\le \frac{2k(1 + (2/3)(n-k)(\gamma+\rho))}{\eta(1 - 4k(n-k)(\gamma+\rho)^2)} + O(1) \\ &\le \frac{2k(5/3)}{\eta(1 - 4/99)} + O(1) \\ &\le \frac{n}{10\eta}\end{aligned}$$

using $\gamma + \rho \le \min\{\frac{1}{n-k}, \frac{1}{99k}\}$, η sufficiently small, and $k \le \frac{n}{100}$ (Assumption 2). By Lemma D.8, $\mathcal{T}^{\varepsilon} = \infty$ if $\frac{n}{10\eta} < \frac{5\varepsilon n}{8\eta}$, which is clearly true for $\varepsilon = \frac{1}{5}$.

We now show that every training point i activates at least one neuron each iteration. By Lemma D.9, this is true if $i \in \mathcal{S}_T$. By Lemma D.11, this is true for $i \in \mathcal{S}_F$ if there is a neuron $j \not\sim i$ such that $\langle \mathbf{w}_j^{(\mathcal{T}_2)}, \mathbf{x}_i \rangle > \eta \frac{2(n-k)(\gamma+\rho)(4k+3)}{0.99(1-\varepsilon)}$. Fix $i \in \mathcal{S}_F$.

First, assume that $\ell(t, \mathbf{x}_i) > 0$ for all $t < \mathcal{T}_2$. By Assumption 2 and Lemma D.3, we know there is $j \in \Gamma_{y_i}$ such that $i \in \mathcal{A}_j^{(1)}$. Using Lemma C.2 and Lemma D.4 we can bound

$$\begin{aligned}\langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i \rangle &\ge \langle \mathbf{w}_j^{(1)}, \mathbf{x}_i \rangle + \eta T_{ij}(1, \mathcal{T}_1) - \eta(\gamma+\rho) G_j^{(i)}(1, \mathcal{T}_1) + \eta(\gamma-\rho) B_j^{(i)}(1, \mathcal{T}_1) \\ &\ge \eta(1 - (\gamma+\rho)(n-k))(\mathcal{T}_1 - 1)\end{aligned}$$

and

$$\langle \mathbf{w}_j^{(\mathcal{T}_2)}, \mathbf{x}_i \rangle \ge \langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i \rangle + \eta T_{ij}(\mathcal{T}_1, \mathcal{T}_2) - \eta(\gamma+\rho) G_j^{(i)}(\mathcal{T}_1, \mathcal{T}_2) + \eta(\gamma-\rho) B_j^{(i)}(\mathcal{T}_1, \mathcal{T}_2).$$

By induction on t we see that

$$G_j(\mathcal{T}_1, t) \le \max_{\mathcal{T}_1 \le t' < t} \left((n-k)(t+1-t') + \frac{(\gamma+\rho) B_j(\mathcal{T}_1, t')}{\gamma - \rho} \right)$$

for $t > \mathcal{T}_1$. The base case $t = \mathcal{T}_1 + 1$ is clear. Suppose the inequality holds for t. Either $G_j(\mathcal{T}_1, t)$ increases by at most $n - k$ or there is some $i' \in \mathcal{S}_T \cap \mathcal{A}_j^{(t+1)}$ with $i' \not\sim j$. By Lemma C.2 and Lemma D.7,

$$\begin{aligned}0 < \langle \mathbf{w}_j^{(t)}, \mathbf{x}_{i'} \rangle &\le \langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_{i'} \rangle - \eta(\gamma-\rho) G_j(\mathcal{T}_1, t+1) + \eta(\gamma+\rho) B_j(\mathcal{T}_1, t) \\ &\le -\eta(\gamma-\rho) G_j(\mathcal{T}_1, t) + \eta(\gamma+\rho) B_j(\mathcal{T}_1, t)\end{aligned}$$

from which the inequality follows. Since $\frac{(\gamma+\rho)B_j(\mathcal{T}_1,t')}{\gamma-\rho} \le 3k(t' - \mathcal{T}_1) < (n-k)(t' - \mathcal{T}_1)$ (using $\rho < \frac{\gamma}{5}$ and $k \le \frac{n}{100}$ from Assumption 2), this maximum occurs at $\tau = \mathcal{T}_1$. This yields

$$\begin{aligned}\langle \mathbf{w}_j^{(\mathcal{T}_2)}, \mathbf{x}_i \rangle &\ge \eta(1 - (\gamma+\rho)(n-k))(\mathcal{T}_2 + 1 - 1)\eta(\gamma-\rho) B_j^{(i)}(\mathcal{T}_1, \mathcal{T}_2) \\ &\ge \eta(1 - (\gamma+\rho)(n-k))\mathcal{T}_2\end{aligned}$$

We want to show this bound is larger than a quantity that is $O(\eta)$. This happens when both of the following hold:

$$O(1) \le (1 - (\gamma+\rho)(n-k))\mathcal{T}_2,$$

which holds when η is sufficiently small.

Now suppose $\ell(\tau, \mathbf{x}_i) = 0$ for some iteration $\tau \leq \mathcal{T}_2$. By Lemma D.7, $\mathcal{T}_1 < \tau$. In this case, we see that

$$\begin{aligned} y_i f(\mathcal{T}_2, \mathbf{x}_i) &= y_i f(\mathcal{T}_2, \mathbf{x}_i) + \sum_{j=1}^{2m} -(-1)^{i+j}[\phi(\langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i\rangle) - \phi(\langle \mathbf{w}_j^{(\tau)}, \mathbf{x}_i\rangle)] \\ &\geq 1 - \eta(\gamma+\rho)G(\tau, \mathcal{T}_2) \\ &\geq 1 - (\gamma+\rho)\frac{2(n-k)}{1-4k(n-k)(\gamma+\rho)^2}\left[\frac{1}{3} + 2k(\gamma+\rho)\right] + O(\eta) \\ &\geq 1 - \frac{2}{1-4/99}\left(\frac{1}{3} + \frac{2}{99}\right) + O(\eta) \geq \frac{1}{4} \end{aligned}$$

where in the third line we use Lemma D.2 and the fourth line we use $\gamma + \rho \leq \min\{\frac{1}{n-k}, \frac{1}{99k}\}$ and η sufficiently small (Assumption 2). Sine this is positive, there is some neuron j with $i \not\sim j$ such that $\phi(\langle \mathbf{w}_j^{(\mathcal{T}_2)}, \mathbf{x}_i\rangle)$ is at least $\frac{1}{m}$ this bound. This is an $\Omega(1)$ lower bound. Since the required condition is $\langle \mathbf{w}_j^{(\mathcal{T}_2)}, \mathbf{x}_i\rangle > O(\eta)$, this can be achieved by taking η sufficiently small. □

D.5 Proof of Lemma 3.5

Lemma D.13 (Lemma 3.5)**.** *Assume Assumption 2 holds. Let $y \in \{-1, 1\}$ chosen uniformly and $\mathbf{x} := (y\sqrt{\gamma}\mathbf{v}, \sqrt{1-\gamma}\mathbf{n})$, where $\mathbf{n} \sim$ Uniform$(\mathcal{S}^{d-1} \cap \operatorname{span}\{\mathbf{v}\}^\perp)$. Suppose that $|\langle \mathbf{n}, \mathbf{n}_\ell\rangle| < \frac{\rho}{1-\gamma}$ for all $l \in [2n]$, then $yf(\mathcal{T}_{end}, \mathbf{x}) > 0$.*

Proof. Following the same steps as in (5) for any $j \in [2m]$

$$\begin{aligned} \langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, \mathbf{x}\rangle &= \langle \mathbf{w}_j^{(1)}, \mathbf{x}_i\rangle + (-1)^j \eta \sum_{\ell=1}^{2n} T_{\ell j}(1, \mathcal{T}_{\text{end}}) y_\ell \langle \mathbf{x}_\ell, \mathbf{x}_i\rangle \\ &= \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i\rangle + (-1)^j y\eta \sum_{\ell=1}^{2n} T_{\ell j}(t_0, t)(-1)^\ell y\beta(\ell)\langle \mathbf{x}_\ell, \mathbf{x}\rangle \\ &= \langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i\rangle + (-1)^j y\eta \sum_{\ell=1}^{2n} T_{\ell j}(t_0, t)\lambda'_{i\ell}, \end{aligned}$$

where $\lambda'_\ell := (-1)^l y\beta(\ell)\langle \mathbf{x}_\ell, \mathbf{x}\rangle = \beta(\ell)\gamma + (1-\gamma)\langle \mathbf{n}_\ell, \mathbf{n}\rangle$. Then as in Lemma C.1

$$\begin{aligned} \gamma - \rho \leq &\lambda'_\ell \leq \gamma + \rho, \text{ if } i \in \mathcal{S}_T, \\ -(\gamma+\rho) \leq &\lambda'_\ell \leq -(\gamma-\rho), \text{ if } i \in \mathcal{S}_F,. \end{aligned}$$

Recall, from Lemma D.4 for any $j \in \Gamma_p$ then $G_j(1, \mathcal{T}_{\text{end}}) \geq G_j(1, \mathcal{T}_1) = \mathcal{T}_1(n-k)$. As a consequence, for $j \in \Gamma_p$ we have

$$\begin{aligned} \langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, \mathbf{x}\rangle &\geq \langle \mathbf{w}_j^{(1)}, \mathbf{x}_i\rangle + \eta G_j(1, \mathcal{T}_{\text{end}})(\gamma-\rho) - \eta B_j(1, \mathcal{T}_{\text{end}})(\gamma+\rho) \\ &\geq O(\eta) + \mathcal{T}_1(n-k)(\gamma-\rho) - \eta B_j(1, \mathcal{T}_{\text{end}})(\gamma+\rho). \end{aligned}$$

For j such that $(-1)^j = y$ then

$$\begin{aligned} \phi(\langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, \mathbf{x}\rangle) &\leq \phi(\langle \mathbf{w}_j^{(1)}, \mathbf{x}_i\rangle - \eta G_j(1, \mathcal{T}_{\text{end}})(\gamma-\rho) + \eta B_j(1, \mathcal{T}_{\text{end}})(\gamma+\rho)) \\ &\leq O(\eta) + \eta B_j(1, \mathcal{T}_{\text{end}})(\gamma+\rho). \end{aligned}$$

As a result

$$\begin{aligned} yf(\mathcal{T}_{\text{end}}, \mathbf{x}) &= \sum_{j\in[2m]} \phi(\langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, \mathbf{x}\rangle) \\ &\geq \sum_{j\in\Gamma_p} \langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, \mathbf{x}\rangle - \sum_{j\,:\,(-1)^j\neq y} \phi(\langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, \mathbf{x}\rangle) \\ &\geq \eta \sum_{j\in\Gamma_p} \mathcal{T}_1(n-k)(\gamma-\rho) - \eta \sum_{j\in[2n]} B_j(1, \mathcal{T}_{\text{end}})(\gamma+\rho) + O(\eta) \\ &\geq 0.99\eta m \mathcal{T}_1(n-k) - \eta B(1, \mathcal{T}_{\text{end}})(\gamma+\rho)) + O(\eta) \end{aligned}$$

using $|\Gamma_p| \geq 0.99$ (Assumption 2). From Lemma D.7

$$\mathcal{T}_1 = \frac{1}{1.03\eta m[1+(\gamma+\rho)(n-k)]} + O(1),$$

furthermore, combining the assumptions $100k < n$, η sufficiently small, and $\gamma+\rho < \frac{1}{n-k}$ with Lemma D.2 we see

$$\begin{aligned} B(1, \mathcal{T}_{\text{end}}) &\leq \frac{2k}{1-4k(n-k)(\gamma+\rho)^2}\left[\frac{1}{\eta} + 2(n-k)(\gamma+\rho)\left(\frac{1}{\eta}\right)\right] + O(n) \\ &\leq \frac{n}{10\eta}. \end{aligned}$$

Here we also use that $\ell(0, \mathbf{x}_i) \leq 1 + m\lambda_w = 1 + O(\eta)$ for all i.

Combining these inequalities it follows that

$$\begin{aligned} yf(\mathcal{T}_{\text{end}}, \mathbf{x}) &\geq \frac{0.99}{1.03} \cdot \frac{n-k}{2}(\gamma-\rho) - \frac{n}{10}(\gamma+\rho) + O(\eta) \\ &\geq \frac{n}{3} - \frac{3n}{25} + O(\eta) > 0, \end{aligned}$$

again using $100k < n$, η sufficiently small, and $\gamma+\rho < \frac{1}{n-k}$. □

D.6 Proof of Theorem 3.1

Theorem D.14 (Theorem 3.1)**.** *Let Assumption 1 hold. With probability at least $1-\delta$ over the randomness of the dataset and network initialization the following hold.*

1. *The training process terminates at an iteration $\mathcal{T}_{end} \leq \frac{Cn}{\eta}$.*
2. *For all $i \in [2n]$ $\ell(\mathcal{T}_{end}, \mathbf{x}_i) = 0$.*
3. *The generalization error satisfies*

$$\mathbb{P}(\operatorname{sgn}(f(\mathcal{T}_{end}, \mathbf{x})) \neq y) \leq C\exp\left(-\frac{d\gamma^2}{32}\right).$$

Proof. Under Assumption 2 Statement 1 and 2 are derived from Lemma D.12. The bound in statement 1 comes from Lemma D.2 applied from iteration 0 to iteration $\mathcal{T}_{\text{end}}$, using $4k(n-k)(\gamma+\rho)^2 < 4/99$ and $\ell(0, \mathbf{x}_i) = 1 + O(\eta)$ for all $i \in [2n]$. Statement 3 comes from Lemma D.13. Finally, under Assumption 1 then by Lemma D.1 Assumption 2 holds probability at least $1-\delta$. □

E Non-benign overfitting

Assumption 3. *Let $\delta \in (0,1)$. For the non-benign overfitting setting we assume the following conditions on the data and model hyperparameters.*

1. $\gamma \leq \frac{1}{6\sqrt{dn}}$

2. $\eta \leq \frac{1}{2mn}$,

3. $\lambda_w \leq \min\{n\gamma/10, \eta/4\}$,

4. $m \geq \log_2(\frac{4n}{\delta})$

5. $d \geq \max\left\{6, 3\rho^{-2}\ln\left(\frac{4n^2}{\delta}\right)\right\}$ *where* $\rho \leq \min\{\frac{1-\gamma}{4n}, \frac{1}{3n} - \gamma\}$.

For convenience, in our analysis we will make two additional assumptions.

Assumption 4. *In addition to the conditions detailed in Assumption 4, assume the following two conditions hold.*

1. *For all* $i \in [2n]$ *there exists a* $j \in [2m]$ *such that* $(-1)^j = y_i$ *and* $i \in \mathcal{A}_j^{(0)}$.

2. *For all* $i, l \in [2n]$, $i \neq l$ $|\langle \mathbf{n}_i, \mathbf{n}_l \rangle| \leq \frac{\rho}{1-\gamma}$.

As demonstrated in the following Lemma, these additional two conditions hold with high probability over the randomness of the initialization and training set.

Lemma E.1. *The additional conditions of Assumption 4 hold with probability at least* $1 - \delta$.

Proof. The additional conditions of Assumption 4 over Assumption 3 are as follows.

1. For all $i \in [2n]$ there exists a $j \in [2m]$ such that $(-1)^j = y_i$ and $i \in \mathcal{A}_j^{(0)}$.

2. For all $i, l \in [2n]$, $i \neq l$ $|\langle \mathbf{n}_i, \mathbf{n}_l \rangle| \leq \frac{\rho}{1-\gamma}$.

Using Lemma B.5, then as long as $m \geq \log_2(\frac{4n}{\delta})$ the probability the first condition does not hold is at most $\delta/2$. Using Lemma B.1, and observing $\frac{\rho}{1-\gamma} > \rho$, then as long as

$$d \geq \max\left\{6, 3\rho^{-2}\ln\left(\frac{4n^2}{\delta}\right)\right\}$$

the probability that the second condition does not hold is also at most $\delta/2$. Using the union bound we conclude that both properties hold with probability at least δ. □

E.1 Proof of Lemma 3.7

Lemma E.2 (Lemma 3.7)**.** *Assume Assumption4 and that* $\ell(t_0, \mathbf{x}_i) \leq a$ *for all* i. *Then*

$$T(t_0, t) \leq \frac{2n}{1 - (2n-1)(\gamma+\rho)}\left(\frac{a}{\eta} + 3m\right).$$

Proof. From Lemma C.5, $\phi(\rho - \gamma) \leq \rho + \gamma$, and the assumption on a,

$$T_i(t_0, t) \leq \frac{a}{\eta} + 3m + (\gamma+\rho)T^{(i)}(t_0, t).$$

If we sum over i, we get

$$T(t_0, t) \leq \frac{2na}{\eta} + 6mn + (2n-1)(\gamma+\rho)T(t_0, t),$$

from which the result follows. □

E.2 Additional lemmas

Lemma E.3. *If Assumption 4 holds, then the training process converges to zero loss.*

Proof. By Lemma E.2, there is an upper bound on the number of updates independent of epoch. This can only happen if there is some epoch after which we make no updates. In turn, this can only happen if every point is either at zero loss or activates no neurons. We prove by induction that every point activates a neuron each epoch $t = 0$. Fix a point i.

At $t = 0$ this is true by initialization. Now suppose it is true at epoch t. There are two cases to consider to show this for epoch $t + 1$:

1. If $\ell(t, \mathbf{x}_i) > 0$, then let j be such that $\phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) > 0$. We can bound

$$\begin{aligned}\phi(\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i \rangle) &\geq \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) + \eta - \eta(\gamma + \rho) H_j^{(i)}(t, t+1) \\ &\geq \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) + \eta[1 - (\gamma + \rho)(2n - 1)] \\ &> 0,\end{aligned}$$

   since $\gamma + \rho < \frac{1}{2n-1}$ (Assumption 4).

2. If $\ell(t, \mathbf{x}_i) = 0$, then

$$\begin{aligned}y_i f(t, \mathbf{x}_i) &= y_i \sum_{j=1}^{2m} (-1)^j \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \\ &\leq \sum_{j \,:\, (-1)^j = y_i} \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle)\end{aligned}$$

   is bounded below by 1. This means that there is some j such that $\phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \geq \frac{1}{m}$. We bound

$$\begin{aligned}\phi(\langle \mathbf{w}_j^{(t+1)}, \mathbf{x}_i \rangle) &\geq \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) - \eta(\gamma + \rho) H_j^{(i)}(t, t+1) \\ &\geq \frac{1}{m} - \eta(\gamma + \rho)(2n - 1) \\ &> 0\end{aligned}$$

   since $\eta < \frac{1}{2mn}$ (Assumption 4). □

Lemma E.4. *Suppose that at epoch τ every point is at zero loss. Then we can bound*

$$\eta T(0, \tau) \geq n + O(\eta).$$

Proof. If $\ell(\tau, \mathbf{x}_i) = 0$ for all i, then $y_i f(\tau, \mathbf{x}_i) \geq 1$ for all i as well. We bound

$$\begin{aligned}y_i f(\tau, \mathbf{x}_i) &\leq \sum_{j \,:\, (-1)^j = y_i} (\phi(\langle \mathbf{w}_j^{(\tau)}, \mathbf{x}_i \rangle) - \phi(\langle \mathbf{w}_j^{(0)}, \mathbf{x}_i \rangle)) + O(\eta) \\ &\leq \sum_{j \,:\, (-1)^j = y_i} \eta[T_{ij}(0, \tau) + (\gamma + \rho) H_j^{(i)}(0, \tau)] + O(\eta) \\ &\leq \eta T_i(0, \tau) + \eta(\gamma + \rho) T^{(i)}(0, \tau) + O(\eta).\end{aligned}$$

If we sum over i, we see that

$$2n \leq \eta[1 + (2n - 1)(\gamma + \rho)] T(0, \tau) \leq 2\eta T(0, \tau) + O(\eta),$$

from which the desired result follows. □

Lemma E.5. *Let $y \in \{-1, 1\}$ chosen uniformly and $\mathbf{x} := (y\sqrt{\gamma}\mathbf{v}, \sqrt{1 - \gamma}\mathbf{n})$, where $\mathbf{n} \sim$ Uniform$(\mathcal{S}^{d-1} \cap \operatorname{span}\{\mathbf{v}\}^\perp)$. If Assumption 4 holds, then*

$$\mathbb{P}(y f(\mathcal{T}_{end}, \mathbf{x}) < 0) \geq \frac{1}{8}$$

Proof. Let $y \sim U(\{-1,1\})$ and consider a clean test point $\mathbf{x} = y(\sqrt{\gamma}\mathbf{v} + \sqrt{1-\gamma}\mathbf{n})$, where $\mathbf{n} \sim U(\mathbb{S}^d \setminus \operatorname{span} \mathbf{v})$. Observe by symmetry of the distributions of both y and $\mathbf{n}$ that $-\mathbf{x}$ is identically distributed to $\mathbf{x}$ and furthermore that the labels of $\mathbf{x}$ and $-\mathbf{x}$ are opposite. As a result, if $y(f(\mathcal{T}_{\text{end}}, \mathbf{x}) - f(\mathcal{T}_{\text{end}}, -\mathbf{x})) < 0$ then at least one of $y(f(\mathcal{T}_{\text{end}}, \mathbf{x}) < 0$ or $-y(f(\mathcal{T}_{\text{end}}, -\mathbf{x}) < 0$, in turn implying at least one of them is misclassified. By construction, $\langle \mathbf{w}_j^{(t)}, \mathbf{x}\rangle > 0$ iff $\langle \mathbf{w}_j^{(t)}, -\mathbf{x}\rangle < 0$, therefore

$$\begin{aligned} y(f(\mathcal{T}_{\text{end}}, \mathbf{x}) - f(\mathcal{T}_{\text{end}}, -\mathbf{x})) &= y\sum_{j=1}^{2m}(-1)^j \left(\phi(\langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, \mathbf{x}\rangle) - \phi(\langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, -\mathbf{x}_i\rangle)\right) \\ &= \sum_{j=1}^{2m} y(-1)^j \langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, \mathbf{x}\rangle. \end{aligned}$$

Unwinding the GD update to a neuron we have

$$\mathbf{w}_j^{(\mathcal{T}_{\text{end}})} = \mathbf{w}_j^{(0)} + \eta \sum_{i=1}^{2n} T_{ij}(0, \mathcal{T}_{\text{end}})(-1)^{i+j}\mathbf{x}_i.$$

Furthermore, as

$$\langle \mathbf{x}_i, \mathbf{x}\rangle = y(-1)^i(\gamma + (1-\gamma)\beta(i)\langle \mathbf{n}_i, \mathbf{n}\rangle))$$

then

$$\begin{aligned} \sum_{j=1}^{2m} y(-1)^j \langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, \mathbf{x}\rangle &= \sum_{j=1}^{2m} y(-1)^j \langle \mathbf{w}_j^{(0)}, \mathbf{x}\rangle + \eta \sum_{j=1}^{2m}\sum_{i=1}^{2n} y(-1)^j T_{ij}(0, \mathcal{T}_{\text{end}})(-1)^{i+j}\langle \mathbf{x}_i, \mathbf{x}\rangle \\ &\le 2m\lambda_w + \eta \sum_{i=1}^{2n} T_i(0, \mathcal{T}_{\text{end}})(\gamma + (1-\gamma)\beta(i)\langle \mathbf{n}_i, \mathbf{n}\rangle)) \\ &= 2m\lambda_w + \eta\left(T(0, \mathcal{T}_{\text{end}})\gamma + \left\langle \mathbf{n}, (1-\gamma)\sum_{i=1}^{2n} T_i(0, \mathcal{T}_{\text{end}})\beta(i)\mathbf{n}_i \right\rangle\right) \\ &\overset{d}{=} 2m\lambda_w + \eta\left(T(0, \mathcal{T}_{\text{end}})\gamma - \|\mathbf{z}\|(1-\gamma)\langle \mathbf{n}, \mathbf{u}\rangle\right), \end{aligned}$$

where the final equality follows from symmetry of the noise distribution, $\mathbf{z} := \sum_{i=1}^{2n} T_i(0, \mathcal{T}_{\text{end}})\mathbf{n}_i$ and $\mathbf{u} = \frac{\mathbf{z}}{\|\mathbf{z}\|}$. Observe

$$\begin{aligned} \|\mathbf{z}\|^2 &= \sum_{i,l=1}^{2n} T_i(0, \mathcal{T}_{\text{end}})T_\ell(0, \mathcal{T}_{\text{end}})\langle \mathbf{n}_i, \mathbf{n}_\ell\rangle \\ &\ge \sum_{i}^{2n} T_i^2(0, \mathcal{T}_{\text{end}}) - \frac{\rho}{1-\gamma}\sum_{i\neq\ell} T_i(0, \mathcal{T}_{\text{end}})T_\ell(0, \mathcal{T}_{\text{end}}) \\ &= \frac{1-\gamma+\rho}{1-\gamma}\sum_{i}^{2n} T_i^2(0, \mathcal{T}_{\text{end}}) - \frac{\rho}{1-\gamma}T^2(0, \mathcal{T}_{\text{end}}) \\ &\ge \left(\frac{1}{2n} - \frac{\rho}{1-\gamma}\right)T^2(0, \mathcal{T}_{\text{end}}). \end{aligned}$$

where the final inequality follows from Jensen's inequality. By assumption $4n\rho < 1-\gamma$, and $10m\lambda_w \le n\gamma \le \frac{\sqrt{n}}{6\sqrt{d}}$, furthermore trivially $(1-\gamma) > 0.8$. Conditioning on the event $\langle \mathbf{n}, \mathbf{u}\rangle > 0$, which holds with probability $1/2$, these inequalities in combination with Lemma E.4 give

$$\begin{aligned} \sum_{j=1}^{2m} y(-1)^j \langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, \mathbf{x}\rangle &\le 2m\lambda_w + n\gamma - \frac{\sqrt{n}}{2}(1-\gamma)\langle \mathbf{n}, \mathbf{u}\rangle \\ &\le \frac{1}{5}\sqrt{\frac{n}{d}} - \frac{4}{10}\sqrt{n}\langle \mathbf{n}, \mathbf{u}\rangle. \end{aligned}$$

Therefore, if $\langle \mathbf{n}, \mathbf{u} \rangle > 0$ then the condition

$$\langle \sqrt{d}\mathbf{n}, \mathbf{u} \rangle \geq \frac{1}{2} \tag{8}$$

implies at least one of $\mathbf{x}$ or $-\mathbf{x}$ is misclassified. Suppose $\mathbf{n} \sim U(\mathbb{S}^d \cap \text{span}(\mathbf{v})^\perp)$ is such that (8) holds. Then as $y \sim U(\{-1, 1\})$ it follows given $\mathbf{n}$ then either $\mathbf{x}$ or $-\mathbf{x}$ are sampled with equal probability and thus the chance of misclassifying is at least $1/2$. As a result the probability of misclassification is at least

$$\frac{1}{4}\mathbb{P}\left(\langle \sqrt{d}\mathbf{n}, \mathbf{u} \rangle \geq \frac{1}{2}\right) \geq \frac{1}{8}$$

as claimed. □

E.3 Proof of Theorem 3.6

Theorem E.6 (Theorem 3.6)**.** *Assume Assumption 3 holds. With probability at least $1 - \delta$ over the randomness of the dataset and network initialization the following hold.*

1. *The training process terminates at an iteration $\mathcal{T}_{end} \leq \frac{Cn}{\eta}$.*
2. *For all $i \in [2n]$ $\ell(\mathcal{T}_{end}, \mathbf{x}_i) = 0$.*
3. *The generalization error satisfies*

$$\mathbb{P}(\text{sgn}(f(\mathcal{T}_{end}, \mathbf{x})) \neq y) \geq \frac{1}{8}.$$

Proof. Under Assumption 4 Statement 1 and 2 come from Lemma E.3. The bound on $\mathcal{T}_{\text{end}}$ comes from Lemma E.2 applied between epoches 0 and $\mathcal{T}_{\text{end}}$, using $\ell(0, \mathbf{x}_i) = 1 + O(\eta)$ for all $i \in [2n]$ and $(\gamma + \rho)(2n - 1) < \frac{2}{3}$. Statement 3 comes from Lemma E.5. We conclude by observing under Assumption 3 that Assumption 4 holds with probability at least $1 - \delta$. □

F Non-overfitting

Assumption 5. *Let $\delta \in (0, 1)$. With $n \geq C$ and $\eta \leq c$ then for no overfitting we assume the following conditions on the data and model hyperparameters.*

1. $k < \frac{n}{100}$,
2. $\frac{3}{n} < \gamma < \frac{1}{36} \min\{k^{-1}, 1\}$,
3. $\lambda_w \leq \frac{\eta\gamma}{4}$
4. $m \geq C \ln\left(\frac{2n}{\delta}\right)$,
5. $d \geq 3\rho^{-2} \ln\left(\frac{4n^2}{\delta}\right)$ *where* $\rho \leq \frac{\gamma}{2}$.

We remark that under these assumptions ρ as given above satisfies the inequality $\rho < \min\left\{\frac{\gamma(n-3k)-2}{n+k}, \frac{\gamma}{5}, \frac{n}{11}\right\}$. For convenience, we will also make two additional assumptions.

Assumption 6. *In addition to the assumptions detailed in Assumption 5, assume the following conditions hold.*

1. $\Gamma = [2m]$.
2. *For all $i, l \in [2n]$ such that $i \neq l$ then $|\langle \mathbf{n}_i, \mathbf{n}_l \rangle| \leq \frac{\rho}{1-\gamma}$.*

As shown in the following lemma, these two additional conditions hold with high probability over the randomness of the initialization and training set.

Lemma F.1. *The extra conditions of Assumption 6 hold with probability at least $1 - \delta$*

Proof. Recall the additional conditions of Assumption 6 over Assumption 5 are as follows,

1. $\Gamma = [2m]$,
2. for all $i, l \in [2n]$, $i \neq l$ $|\langle \mathbf{n}_i, \mathbf{n}_l \rangle| \leq \frac{\rho}{1-\gamma}$.

Using Lemma B.3, then for sufficiently large n there exists a constant c such that the probability the first condition does not hold is at most $m \exp(-cn)$. Alternatively, setting $\delta \geq 2m \exp(-cn)$ and rearranging, as long as $m \geq C \ln\left(\frac{2n}{\delta}\right)$ then the probability the first condition does not hold is at most $\delta/2$.

Using Lemma B.1, observing $\frac{\rho}{1-\gamma} > \rho$ and under the conditions of the lemma that $3\rho^{-2} \ln\left(4n^2\right) > 6$, then as long as

$$d \geq 3\rho^{-2} \ln\left(\frac{4n^2}{\delta}\right)$$

the probability that the second condition does not hold is also at most $\delta/2$. Using the union bound we therefore conclude that both properties hold with probability at least δ. □

F.1 Proof of Lemma 3.9

Lemma F.2 (Lemma 3.9)**.** *Suppose Assumption 6 holds. Consider an arbitrary $j \in [2m]$ and iteration t satisfying $2 \leq t < \mathcal{T}_0$. Then $i \in \mathcal{A}_j^{(t)}$ iff $i \sim j$.*

Proof. First we establish at iteration $t = 1$ that for all $i \in \mathcal{S}_T$, $i \in \mathcal{A}_j^{(t)}$ iff $i \sim j$. The argument here is similar to that of Lemma D.3. Suppose $i \sim j$ and $i \in \mathcal{S}_T$. Recall from definition of Γ_p that $G_j^{(i)}(0,1)(\gamma - \rho) - B_j^{(i)}(0,1)(\gamma + \rho) \geq \frac{2\lambda_w}{\eta}$. By Assumption 6, all neurons are in Γ. Using Lemma C.2

$$\begin{aligned}
\langle \mathbf{w}_j^{(1)}, \mathbf{x}_i \rangle &\geq \langle \mathbf{w}_j^{(0)}, \mathbf{x}_i \rangle + \eta \left(T_{ij}(0,1) + G_j^{(i)}(0,1)(\gamma - \rho) - B_j^{(i)}(0,1)(\gamma + \rho) \right) \\
&> \langle \mathbf{w}_j^{(0)}, \mathbf{x}_i \rangle + \eta \left(G_j(0,1)(\gamma - \rho) - B_j(0,1)(\gamma + \rho) \right) \\
&\geq \langle \mathbf{w}_j^{(0)}, \mathbf{x}_i \rangle + 2\lambda_w \\
&> \lambda_w.
\end{aligned}$$

On the other hand, if $i \not\sim j$ then again from Lemma C.2

$$\begin{aligned}
\langle \mathbf{w}_j^{(1)}, \mathbf{x}_i \rangle &\leq \langle \mathbf{w}_j^{(0)}, \mathbf{x}_i \rangle - \eta \left(T_{ij}(0,1) + G_j^{(i)}(0,1)(\gamma - \rho) - B_j^{(i)}(0,1)(\gamma + \rho) \right) \\
&\leq \langle \mathbf{w}_j^{(0)}, \mathbf{x}_i \rangle - \eta \left(G_j(0,1)(\gamma - \rho) - B_j(0,1)(\gamma + \rho) \right) \\
&\leq -\lambda_w.
\end{aligned}$$

Now we consider $t = 2$. If $i \in \mathcal{S}_T$ and $i \sim j$ then

$$\begin{aligned}
\langle \mathbf{w}_j^{(2)}, \mathbf{x}_i \rangle &\geq \langle \mathbf{w}_j^{(1)}, \mathbf{x}_i \rangle + \eta \left(T_{ij}(1,2) + G_j^{(i)}(1,2)(\gamma - \rho) - B_j^{(i)}(1,2)(\gamma + \rho) \right) \\
&> \eta \left(1 + (n - k - 1)(\gamma - \rho) - 2k(\gamma + \rho) \right)
\end{aligned}$$

whereas if $i \in \mathcal{S}_T$ and $i \not\sim j$ then

$$\begin{aligned}
\langle \mathbf{w}_j^{(2)}, \mathbf{x}_i \rangle &\leq \langle \mathbf{w}_j^{(1)}, \mathbf{x}_i \rangle - \eta \left(T_{ij}(1,2) + G_j^{(i)}(1,2)(\gamma - \rho) - B_j^{(i)}(1,2)(\gamma + \rho) \right) \\
&\leq -\eta((n - k)(\gamma - \rho) - 2k(\gamma + \rho)).
\end{aligned}$$

By assumption $\gamma > \frac{2+(n+k)\rho}{n-3k} > \frac{(n+k)\rho}{n-3k}$ and therefore for $i \in \mathcal{S}_T$ then $i \in \mathcal{A}_j^{(2)}$ iff $i \sim j$. Again using Lemma C.2, for $i \in \mathcal{S}_F$ and $i \sim j$

$$\begin{aligned}
\langle \mathbf{w}_j^{(1)}, \mathbf{x}_i \rangle &\geq \langle \mathbf{w}_j^{(0)}, \mathbf{x}_i \rangle - \eta \left(T_{ij}(0,1) - G_j^{(i)}(0,1)(\gamma - \rho) + B_j^{(i)}(0,1)(\gamma + \rho) \right) \\
&> -\lambda_w - \eta \left(1 - \frac{2\lambda_w}{\eta} \right) \\
&> -\eta,
\end{aligned}$$

and for $i \in \mathcal{S}_F$ and $i \not\sim j$

$$\begin{aligned}\langle \mathbf{w}_j^{(1)}, \mathbf{x}_i \rangle &\leq \langle \mathbf{w}_j^{(0)}, \mathbf{x}_i \rangle + \eta \left(T_{ij}(0,1) - G_j^{(i)}(0,1)(\gamma - \rho) + B_j^{(i)}(0,1)(\gamma + \rho) \right) \\ &< \lambda_w + \eta \left(1 - \frac{2\lambda_w}{\eta} \right) \\ &< \eta.\end{aligned}$$

Therefore, as $\gamma > \frac{2+(n+k)\rho}{n-3k}$ then for $i \in \mathcal{S}_F$ and $i \sim j$

$$\begin{aligned}\langle \mathbf{w}_j^{(2)}, \mathbf{x}_i \rangle &\geq \langle \mathbf{w}_j^{(1)}, \mathbf{x}_i \rangle - \eta \left(T_{ij}(1,2) - G_j^{(i)}(1,2)(\gamma - \rho) + B_j^{(i)}(1,2)(\gamma + \rho) \right) \\ &> -\eta + \eta \left((n-k)(\gamma - \rho) - 2k(\gamma + \rho) - 1 \right) \\ &> 0\end{aligned}$$

and for $i \in \mathcal{S}_F$ and $i \not\sim j$

$$\begin{aligned}\langle \mathbf{w}_j^{(2)}, \mathbf{x}_i \rangle &\leq \langle \mathbf{w}_j^{(1)}, \mathbf{x}_i \rangle + \eta \left(T_{ij}(1,2) - G_j^{(i)}(1,2)(\gamma - \rho) + B_j^{(i)}(1,2)(\gamma + \rho) \right) \\ &< \eta - \eta \left((n-k)(\gamma - \rho) - 2k(\gamma + \rho) - 1 \right) \\ &< 0.\end{aligned}$$

With the base case established we proceed by induction to prove if $i \in \mathcal{A}_j^{(t-1)}$ iff $i \sim j$ then $i \in \mathcal{A}_j^{(t)}$ iff $i \sim j$. By the assumptions on γ, the induction hypothesis and again using Lemma C.2, then for $i \sim j$

$$\begin{aligned}\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle &\geq \langle \mathbf{w}_j^{(t-1)}, \mathbf{x}_i \rangle - \eta \left(T_{ij}(t-1,t) - G_j^{(i)}(t-1,t)(\gamma - \rho) + B_j^{(i)}(t-1,t)(\gamma + \rho) \right) \\ &> \eta \left((n-k)(\gamma - \rho) - k(\gamma + \rho) - 1 \right) \\ &> 0\end{aligned}$$

and for $i \not\sim j$

$$\begin{aligned}\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle &\leq \langle \mathbf{w}_j^{(t-1)}, \mathbf{x}_i \rangle + \eta \left(T_{ij}(t-1,t) - G_j^{(i)}(t-1,t)(\gamma - \rho) + B_j^{(i)}(t-1,t)(\gamma + \rho) \right) \\ &< -\eta((n-k)(\gamma - \rho) - k(\gamma + \rho) - 1) \\ &< 0.\end{aligned}$$

Therefore for an epoch t satisfying $2 \leq t \leq \mathcal{T}_1$ $i \in \mathcal{A}_j^{(t-1)}$ iff $i \sim j$. □

F.2 Proof of Lemma 3.10

Lemma F.3 (Lemma 3.10)**.** *Assume Assumption 6. Then there is an epoch $\mathcal{T}_1 < \mathcal{T}_0$ such that*

$$\begin{aligned}\langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i \rangle &\leq \frac{\gamma(n-2k) + \rho n - 1 + (\gamma - \rho)}{m(1 + \gamma(n-2k) + \rho n - (\gamma - \rho))} + O(\eta) \text{ if } i \in \mathcal{S}_F, i \sim j \\ \langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i \rangle &\geq \frac{1 + \gamma(n-2k) - \rho n - (\gamma - \rho)}{m(1 + \gamma(n-2k) + \rho n - (\gamma - \rho))} + O(\eta) \text{ if } i \in \mathcal{S}_T, i \sim j \\ \langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i \rangle &\leq -\frac{\gamma(n-2k) - \rho n}{m(1 + \gamma(n-2k) + \rho n - (\gamma - \rho))} + O(\eta) \text{ if } i \not\sim j\end{aligned}$$

and for $i \in \mathcal{S}_T$

$$\ell(\mathcal{T}_1, \mathbf{x}_i) \leq \frac{2\rho n}{1 + \gamma(n-2k) + \rho n - (\gamma - \rho)} + O(\eta).$$

Furthermore,

$$\mathcal{T}_1 = \frac{1}{\eta m(1 + \gamma(n-2k) + \rho n - (\gamma - \rho))} + O(1).$$

Proof. Let $t < \mathcal{T}_0$. By Lemma F.2 and Lemma C.2 we can bound, for $i \in \mathcal{S}_F$ and $i \sim j$,

$$\begin{aligned}\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle &\leq \langle \mathbf{w}_j^{(2)}, \mathbf{x}_i \rangle - \eta T_{ij}(2,t) - \eta(\gamma-\rho)B_j^{(i)}(2,t) + \eta(\gamma+\rho)G_j(2,t) \\ &= \eta t((\gamma+\rho)(n-k) - (\gamma-\rho)(k-1) - 1) + O(\eta) \\ &= \eta t(\gamma(n-2k) + \rho n - 1 + (\gamma-\rho)) + O(\eta).\end{aligned}$$

Similarly, for $i \in \mathcal{S}_F$, $i \not\sim j$,

$$\begin{aligned}\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle &\leq \langle \mathbf{w}_j^{(2)}, \mathbf{x}_i \rangle + \eta T_{ij}(2,t) + \eta(\gamma+\rho)B_j^{(i)}(2,t) - \eta(\gamma-\rho)G_j(2,t) \\ &= -\eta t((\gamma-\rho)(n-k) - (\gamma+\rho)k) + O(\eta) \\ &= -\eta t(\gamma(n-2k) - \rho n) + O(\eta).\end{aligned}$$

For $i \in \mathcal{S}_T$, $i \sim j$,

$$\begin{aligned}\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle &\leq \langle \mathbf{w}_j^{(2)}, \mathbf{x}_i \rangle + \eta T_{ij}(2,t) + \eta(\gamma+\rho)G_j^{(i)}(2,t) - \eta(\gamma-\rho)G_j(2,t) \\ &= \eta t(1 + (\gamma+\rho)(n-k-1) - (\gamma-\rho)k) + O(\eta) \\ &= \eta t(1 + \gamma(n-2k) + \rho n - (\gamma+\rho)) + O(\eta)\end{aligned}$$

and

$$\begin{aligned}\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle &\geq \langle \mathbf{w}_j^{(2)}, \mathbf{x}_i \rangle + \eta T_{ij}(2,t) + \eta(\gamma-\rho)G_j^{(i)}(2,t) - \eta(\gamma+\rho)G_j(2,t) \\ &= \eta t(1 + (\gamma-\rho)(n-k-1) - (\gamma+\rho)k) + O(\eta) \\ &= \eta t(1 + \gamma(n-2k) - \rho n - (\gamma-\rho)) + O(\eta).\end{aligned}$$

Lastly, for $i \in \mathcal{S}_T$, $i \not\sim j$,

$$\begin{aligned}\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle &\leq \langle \mathbf{w}_j^{(2)}, \mathbf{x}_i \rangle + \eta T_{ij}(2,t) - \eta(\gamma-\rho)G_j^{(i)}(2,t) + \eta(\gamma+\rho)G_j(2,t) \\ &= -\eta t((\gamma-\rho)(n-k) - (\gamma+\rho)k) + O(\eta) \\ &= -\eta t(\gamma(n-2k) - \rho n) + O(\eta)\end{aligned}$$

Therefore, for $i \in \mathcal{S}_T$,

$$\begin{aligned}f(t, \mathbf{x}_i) &= y_i \sum_{j=1}^{2m} (-1)^j \phi(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \\ &= \sum_{j \sim i} \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle\end{aligned}$$

from which we conclude

$$\eta m t(1+\gamma(n-2k)-\rho n-(\gamma-\rho)) + O(\eta) \leq f(t, \mathbf{x}_i) \leq \eta m t(1+\gamma(n-2k)+\rho n-(\gamma-\rho)) + O(\eta).$$

Therefore, as long as

$$\eta m t(1 + \gamma(n-2k) + \rho n - (\gamma-\rho)) + O(\eta) < 1, \tag{9}$$

then $\ell(t, \mathbf{x}_i) > 0$. Let $\mathcal{T}_1$ be the largest value of t satisfying (9) and $t < \mathcal{T}_0$. We see that

$$\mathcal{T}_1 = \frac{1}{\eta m(1 + \gamma(n-2k) + \rho n - (\gamma-\rho))} + O(1).$$

From this, the desired bounds follow. □

F.3 Proof of Lemma 3.11

Lemma F.4 (Lemma 3.11)**.** *Let Assumption 6 hold. Suppose there is a time t_0 so that:*

a. $\ell(t_0, \mathbf{x}_i) \leq a$ for all $i \in \mathcal{S}_T$,

b. $\phi(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle) \leq b$ for all $i \in \mathcal{S}_F$ and $i \sim j$,

c. For all iterations τ satisfying $t_0 \leq \tau \leq t$ it holds that $i \in \mathcal{A}_j^{(\tau)}$ only if $i \sim j$,

d. For all iterations τ satisfying $t_0 \le \tau \le t$, $i \in \mathcal{A}_j^{(\tau)}$ if $i \sim j$ and $i \in \mathcal{S}_T$.

We can bound

$$B_j(t_0, \tau) \le k\left(\frac{3b}{2\eta} + \frac{2a}{\eta m}\right)$$
$$\sum_{j \sim s} G_j(t_0, t) \le \frac{1}{\gamma}\left(\frac{3a}{2\eta} + \frac{mb}{\eta}\right).$$

Proof. Fix a neuron j. Using Lemma C.3 and assumption (b) we bound for $t < \tau \le t$, $i \in \mathcal{S}_F$, and $i \sim j$

$$\phi(\langle \mathbf{w}_j^{(\tau-1)}, \mathbf{x}_i \rangle) \le b - \eta(T_{ij}(t_0, \tau-1) - (\gamma+\rho)G_j(t_0, \tau-1) - 1).$$

There are two possibilities. If this bound is negative, then $T_{ij}(t_0, \tau) = T_{ij}(t_0, \tau-1)$. Otherwise,

$$\begin{aligned} T_{ij}(t_0, \tau) &\le T_{ij}(t_0, \tau-1) + 1 \\ &\le \frac{b}{\eta} + (\gamma+\rho)G_j(t_0, \tau-1) + 2 \\ &\le \frac{b}{\eta} + (\gamma+\rho)G_j(t_0, \tau) + 2 \end{aligned}$$

This yields, by induction,

$$T_{ij}(t_0, \tau) \le \frac{b}{\eta} + (\gamma+\rho)G_j(t_0, \tau) + 2.$$

We then sum over all $i \in \mathcal{S}_F$ such that $i \sim j$. By assumption (c),

$$\sum_{\substack{i \in \mathcal{S}_F \\ i \sim j}} T_{ij}(t_0, \tau) = B_j(t_0, \tau).$$

This yields

$$B_j(t_0, \tau) \le \frac{kb}{\eta} + k(\gamma+\rho)B_j(t_0, \tau) + 2k$$

which, by our assumption, is equivalent to

$$B_j(t_0, \tau) \le \frac{kb}{\eta} + \frac{k(\gamma+\rho)}{m}\sum_{\ell \sim j} G_\ell(t_0, t) \tag{10}$$

since the number of clean updates on two neurons ℓ and j with $\ell \sim j$ is the same by assumption (c) and (d).

Fix $s \in [2m]$. We now bound $\sum_{j \sim s} G_j(t_0, t)$. This time, we use Lemma C.2 to bound for $i \in \mathcal{S}_T$ and $i \sim j$

$$\langle \mathbf{w}_j^{(\tau-1)}, \mathbf{x}_i \rangle \ge \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle + \eta(T_{ij}(t_0, \tau-1) + (\gamma-\rho)G_j^{(i)}(t_0, \tau-1) - (\gamma+\rho)B_j(t_0, \tau-1)).$$

Using assumption (c) and (d),

$$\begin{aligned} y_i f(\tau-1, \mathbf{x}_i) &= \sum_{j \sim i} \phi(\langle \mathbf{w}_j^{(\tau-1)}, \mathbf{x}_i \rangle) \\ &\ge \sum_{j \sim i} \Big(\langle \mathbf{w}_j^{(t_0)}, \mathbf{x}_i \rangle + \eta T_{ij}(t_0, \tau-1) \\ &\qquad + \eta(\gamma-\rho)G_j^{(i)}(t_0, \tau-1) - \eta(\gamma+\rho)B_j(t_0, \tau-1) \Big) \\ &\ge (1-a) + \eta(1-(\gamma-\rho))T_i(t_0, \tau-1) \\ &\qquad + \left(\sum_{j \sim i} ((\gamma-\rho)G_j(t_0, \tau-1) - (\gamma+\rho)B_j(t_0, \tau-1)) \right). \end{aligned}$$

Either $T_i(t_0, \tau) = T_i(t_0, \tau - 1)$ or $\ell(\tau, \mathbf{x}_i) > 0$. If the latter holds, then

$$\eta((1 - (\gamma - \rho))T_i(t_0, \tau - 1) + \sum_{j \sim i} ((\gamma - \rho)G_j(t_0, \tau - 1) - (\gamma + \rho)B_j(t_0, \tau - 1))) < a.$$

Suppose $\tau' \leq \tau$ is the first iteration before τ such that $G_j(\tau', \tau) = 0$. Let $i \in \mathcal{S}_T$, $i \sim s$ be a point that makes an update at iteration $\tau' - 1$. Using the above bound, we see

$$\eta((1 - (\gamma - \rho))T_i(t_0, \tau' - 1) + \sum_{j \sim i} ((\gamma - \rho)G_j(t_0, \tau' - 1) - (\gamma + \rho)B_j(t_0, \tau' - 1))) < a.$$

which implies

$$\sum_{j \sim i} ((\gamma - \rho)G_j(t_0, \tau' - 1) - (\gamma + \rho)B_j(t_0, \tau' - 1)) < \frac{a}{\eta}$$

and

$$\sum_{j \sim s} ((\gamma - \rho)G_j(t_0, \tau') - (\gamma + \rho)B_j(t_0, \tau')) < \frac{a}{\eta} - 2(n - k)(\gamma - \rho).$$

By definition of τ', we conclude

$$\sum_{j \sim s} ((\gamma - \rho)G_j(t_0, \tau) - (\gamma + \rho)B_j(t_0, \tau)) < \frac{a}{\eta} - 2(n - k)(\gamma - \rho).$$

From this we get the bound

$$\sum_{j \sim s} G_j(t_0, t) \leq \frac{1}{\gamma - \rho} \left(\frac{a}{\eta} + (\gamma + \rho) \sum_{j \sim s} B_j(t_0, t) \right) + O(1). \tag{11}$$

We combine (10) with (11) summed over $j \sim s$ to see

$$\begin{aligned} \sum_{j \sim s} G_j(t_0, t) &\leq \frac{1}{\gamma - \rho} \left(\frac{a}{\eta} + (\gamma + \rho) \left(\frac{kmb}{\eta} + k(\gamma + \rho) \sum_{j \sim s} G_j(t_0, t) \right) \right) + O(1) \\ &\leq \frac{1}{\gamma - \rho - k(\gamma + \rho)^2} \left(\frac{a}{\eta} + \frac{kmb(\gamma + \rho)}{\eta} \right) + O(1) \end{aligned}$$

and

$$B_j(t_0, \tau) \leq \frac{kb}{\eta} + \frac{k(\gamma + \rho)}{m(\gamma - \rho - k(\gamma + \rho)^2)} \left(\frac{a}{\eta} + \frac{kmb(\gamma + \rho)}{\eta} \right) + O(1).$$

Using $\rho \leq \frac{\gamma}{5}$, η sufficiently small, and $\gamma \leq \frac{1}{36k}$ (Assumption 6) these bounds simplify to

$$\begin{aligned} B_j(t_0, \tau) &\leq k \left(\frac{3b}{2\eta} + \frac{2a}{\eta m} \right) \\ \sum_{j \sim s} G_j(t_0, t) &\leq \frac{1}{\gamma} \left(\frac{3a}{2\eta} + \frac{mb}{\eta} \right). \end{aligned}$$

□

F.4 Late training

Lemma F.5. *Under Assumption 6, the training process terminates at an iteration $\mathcal{T}_{end}$ satisfying*

$$\ell(\mathcal{T}_{end}, \mathbf{x}_i) = 0$$

for all $i \in \mathcal{S}_T$ and

$$\phi(\langle \mathbf{w}_j^{(\mathcal{T}_{end})}, \mathbf{x}_i \rangle) = 0$$

for all $i \in \mathcal{S}_F$ and $j \in [2m]$.

Proof. By Lemma F.3 and Lemma F.4, we converge provided that the assumptions of the lemmas are satisfied. Since Lemma F.3 requires only Assumption 6, we turn our attention towards Lemma F.4. This has more delicate conditions. Let $t_0 = \mathcal{T}_1$. We first see that the conditions on a and b are satisfied if we take

$$a = \frac{2\rho n}{1+\gamma(n-2k)+\rho n-(\gamma-\rho)} + O(\eta)$$
$$b = \frac{\gamma(n-2k)+\rho n-1+(\gamma-\rho)}{m(1+\gamma(n-2k)+\rho n-(\gamma-\rho))} + O(\eta).$$

using Lemma F.3. Next, using Lemma C.2 we see by induction on $t \geq \mathcal{T}_1$ that if

$$B_j(\mathcal{T}_1, t) < \frac{\gamma(n-2k)-\rho n}{\eta m(\gamma+\rho)(1+\gamma(n-2k)+\rho n-(\gamma-\rho))} + O(1)$$

then for $i \not\sim j$ and $\mathcal{T}_1 \leq \tau \leq t$,

$$\begin{aligned}\langle \mathbf{w}_j^{(\tau)}, \mathbf{x}_i\rangle &\leq \langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i\rangle + \eta(\gamma+\rho)B_j(\mathcal{T}_1,\tau)\\ &\leq \langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i\rangle + \eta(\gamma+\rho)B_j(\mathcal{T}_1,t)\\ &< 0\end{aligned}$$

and for $i \in \mathcal{S}_T$, $i \sim j$, and $\mathcal{T}_1 \leq \tau \leq t$,

$$\begin{aligned}\langle \mathbf{w}_j^{(\tau)}, \mathbf{x}_i\rangle &\geq \langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i\rangle - \eta(\gamma+\rho)B_j(\mathcal{T}_1,\tau)\\ &\geq \langle \mathbf{w}_j^{(\mathcal{T}_1)}, \mathbf{x}_i\rangle - \eta(\gamma+\rho)B_j(\mathcal{T}_1,t)\\ &\geq \frac{1-(\gamma-\rho)}{m(1+\gamma(n-2k)+\rho n-(\gamma-\rho))} + O(\eta)\\ &> 0\end{aligned}$$

for η sufficiently small. So we converge to the desired steady state so long as

$$k\left(\frac{3b}{2\eta}+\frac{2a}{\eta m}\right) < \frac{\gamma(n-2k)-\rho n}{\eta m(\gamma+\rho)(1+\gamma(n-2k)+\rho n-(\gamma-\rho))} + O(1)$$

which is equivalent to

$$k(\gamma+\rho)\left(\frac{3}{2}(\gamma(n-2k)-1+(\gamma-\rho))+\frac{11}{2}\rho n\right) < \gamma(n-2k)-\rho n+O(\eta).$$

This is true by Assumption 6, as

$$\begin{aligned}k(\gamma+\rho)\left(\frac{3}{2}(\gamma(n-2k)-1+(\gamma-\rho))+\frac{11}{2}\rho n\right) &\leq \frac{1}{30}\left(\frac{3}{2}\gamma(n-2k)-\frac{3}{2}+\frac{1}{36}+\frac{1}{2}\right)\\ &\leq \frac{1}{20}\gamma(n-2k)-\frac{1}{2}\\ &< \gamma(n-2k)-\rho n\end{aligned}$$

using $\gamma \leq \min\left\{\frac{1}{36k}, \frac{1}{36}\right\}$ and $\rho \leq \min\left\{\frac{\gamma}{5}, \frac{n}{11}\right\}$. □

Lemma F.6. *Assume Assumption 6 holds. Let $y \in \{-1,1\}$ chosen uniformly and $\mathbf{x} := (y\sqrt{\gamma}\mathbf{v}, \sqrt{1-\gamma}\mathbf{n})$, where $\mathbf{n} \sim$ Uniform$(\mathcal{S}^{d-1} \cap \text{span}\{\mathbf{v}\}^\perp)$. Suppose that $|\langle \mathbf{n}, \mathbf{n}_\ell\rangle| < \frac{\rho}{1-\gamma}$ for all $l \in [2n]$, then $yf(\mathcal{T}_{end}, \mathbf{x}) > 0$.*

Proof. We proceed similarly to Lemma D.13.

Following the same steps as in (5) for any $j \in [2m]$

$$\begin{aligned}\langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, \mathbf{x}\rangle &= \langle \mathbf{w}_j^{(2)}, \mathbf{x}_i\rangle + (-1)^j\eta\sum_{\ell=1}^{2n} T_{\ell j}(1,\mathcal{T}_{\text{end}})y_\ell\langle \mathbf{x}_\ell, \mathbf{x}_i\rangle\\ &= \langle \mathbf{w}_j^{(2)}, \mathbf{x}_i\rangle + (-1)^j y\eta\sum_{\ell=1}^{2n} T_{\ell j}(2,t)(-1)^\ell y\beta(\ell)\langle \mathbf{x}_\ell, \mathbf{x}\rangle\\ &= \langle \mathbf{w}_j^{(2)}, \mathbf{x}_i\rangle + (-1)^j y\eta\sum_{\ell=1}^{2n} T_{\ell j}(2,t)\lambda'_{i\ell},\end{aligned}$$

where $\lambda'_\ell := (-1)^l y\beta(\ell)\langle \mathbf{x}_\ell, \mathbf{x}\rangle = \beta(\ell)\gamma + (1-\gamma)\langle \mathbf{n}_\ell, \mathbf{n}\rangle$. Then as in Lemma C.1

$$\begin{aligned}\gamma - \rho \leq &\lambda'_\ell \leq \gamma + \rho, \text{ if } i \in \mathcal{S}_T,\\ -(\gamma+\rho) \leq &\lambda'_\ell \leq -(\gamma-\rho), \text{ if } i \in \mathcal{S}_F,.\end{aligned}$$

Recall, from Lemma F.2 for any $j \in [2m]$ then $G_j(2, \mathcal{T}_{\text{end}}) \geq G_j(2, \mathcal{T}_1) = (\mathcal{T}_1 - 2)(n-k)$. As a consequence, for $j \in \Gamma_p$ we have

$$\begin{aligned}\langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, \mathbf{x}\rangle &\geq \langle \mathbf{w}_j^{(2)}, \mathbf{x}_i\rangle + \eta G_j(2, \mathcal{T}_{\text{end}})(\gamma-\rho) - \eta B_j(2, \mathcal{T}_{\text{end}})(\gamma+\rho)\\ &\geq O(\eta) + \mathcal{T}_1(n-k)(\gamma-\rho) - \eta B_j(2, \mathcal{T}_{\text{end}})(\gamma+\rho).\end{aligned}$$

For j such that $(-1)^j = y$ then

$$\begin{aligned}\phi(\langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, \mathbf{x}\rangle) &\leq \phi(\langle \mathbf{w}_j^{(2)}, \mathbf{x}_i\rangle - \eta G_j(2, \mathcal{T}_{\text{end}})(\gamma-\rho) + \eta B_j(2, \mathcal{T}_{\text{end}})(\gamma+\rho))\\ &\leq O(\eta) + \eta B_j(2, \mathcal{T}_{\text{end}})(\gamma+\rho).\end{aligned}$$

As a result

$$\begin{aligned}y f(\mathcal{T}_{\text{end}}, \mathbf{x}) &= \sum_{j\in[2m]} \phi(\langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, \mathbf{x}\rangle)\\ &\geq \sum_{j\,:\,(-1)^j = y} \langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, \mathbf{x}\rangle - \sum_{j\,:\,(-1)^j \neq y} \phi(\langle \mathbf{w}_j^{(\mathcal{T}_{\text{end}})}, \mathbf{x}\rangle)\\ &\geq \eta \sum_{j\,:\,(-1)^j = y} (\mathcal{T}_1 - 2)(n-k)(\gamma-\rho) - \eta \sum_{j\in[2n]} B_j(2, \mathcal{T}_{\text{end}})(\gamma+\rho) + O(\eta)\\ &\geq \eta m \mathcal{T}_1 (n-k)(\gamma-\rho) - \eta B(2, \mathcal{T}_{\text{end}})(\gamma+\rho)) + O(\eta)\end{aligned}$$

We decompose $B(2, \mathcal{T}_{\text{end}}) = B(2, \mathcal{T}_1) + B(\mathcal{T}_1, \mathcal{T}_{\text{end}})$. From Lemma F.2,

$$B(2, \mathcal{T}_1) = 2km(\mathcal{T}_1 - 2) = 2km\mathcal{T}_1 + O(1).$$

From Lemma F.3 and Lemma F.4, using the assumptions $\rho \leq \frac{n}{11}$, η sufficiently small, and $\gamma \leq \min\{\frac{k}{36}, \frac{1}{36}\}$,

$$\begin{aligned}B(\mathcal{T}_1, \mathcal{T}_{\text{end}}) &\leq 2mk\left(\frac{3b}{2\eta} + \frac{2a}{\eta m}\right)\\ &\leq 2mk\mathcal{T}_1\left(\frac{3(\gamma(n-2k) + \rho n - 1 + (\gamma-\rho))}{2} + 4\rho n\right) + O(1)\\ &\leq 2m\mathcal{T}_1\left(\frac{n-2k}{24} + \frac{1}{22} - \frac{1}{2} + \frac{1}{72} + \frac{4}{11}\right) + O(1)\\ &\leq \frac{m\mathcal{T}_1(n-k)}{12}.\end{aligned}$$

Using the assumption that $k \leq \frac{n}{100}$ and η is sufficiently small we see that

$$B(2, \mathcal{T}_{\text{end}}) \leq \frac{m\mathcal{T}_1(n-k)}{9}.$$

Since

$$y f(\mathcal{T}_{\text{end}}, \mathbf{x}) \geq \eta m \mathcal{T}_1 (n-k)((\gamma-\rho) - (\gamma+\rho)/9) + O(\eta),$$

$y f(\mathcal{T}_{\text{end}}, \mathbf{x})$ is positive provided $(\gamma-\rho) - (\gamma+\rho)/9$ is positive and η is sufficiently small. Both are guaranteed by Assumption 6, so the point is correctly classified. □

F.5 Proof of Theorem 3.8

Theorem F.7 (Theorem 3.8)**.** *Assume Assumption 5 holds. With probability at least $1-\delta$ over the randomness of the dataset and network initialization we have the following.*

1. *The training process terminates at an iteration $\mathcal{T}_{end} \leq \frac{Cn}{\eta}$.*

2. *For all* $i \in \mathcal{S}_T$ *then* $\ell(\mathcal{T}_{end}, \mathbf{x}_i) = 0$ *while for all* $i \in \mathcal{S}_F$ $\ell(\mathcal{T}_{end}, \mathbf{x}_i) = 1$.

3. *The generalization error satisfies*

$$\mathbb{P}(\mathrm{sgn}(f(\mathcal{T}_{end}, \mathbf{x})) \neq y) \leq C \exp\left(-\frac{d}{2n^2}\right).$$

Proof. Under Assumption 6, statements (1) and (2) follow from Lemma F.5. The bound on $\mathcal{T}_{\text{end}}$ follows from Lemma F.4 applied at $t_0 = 2$:

$$B(2, \mathcal{T}_{\text{end}}) \leq k\left(\frac{4}{\eta m}\right) + O(\eta)$$
$$B(2, \mathcal{T}_{\text{end}}) \leq \frac{2}{\gamma}\left(\frac{3}{2\eta}\right) + O(\eta)$$

with $\gamma > \frac{3}{n}$. Statement (3) is derived from Lemma F.6. Finally, Lemma F.1 implies that under Assumption 5 then Assumption 6 holds with probability at least $1 - \delta$. □

G Numerical simulations

Reproducibility statement: the code used to generate the following figures can be found at `https://anonymous.4open.science/r/benign_overfitting-4A4C/BO_experiments.ipynb`.

To investigate our theory we train two-layer neural networks with ReLU activations using full-batch gradient descent and a fixed step size. We train on the synthetic binary classification dataset, detailed in Section B.2, that we have studied throughout the paper. Finally we train using both hinge and logistic loss.

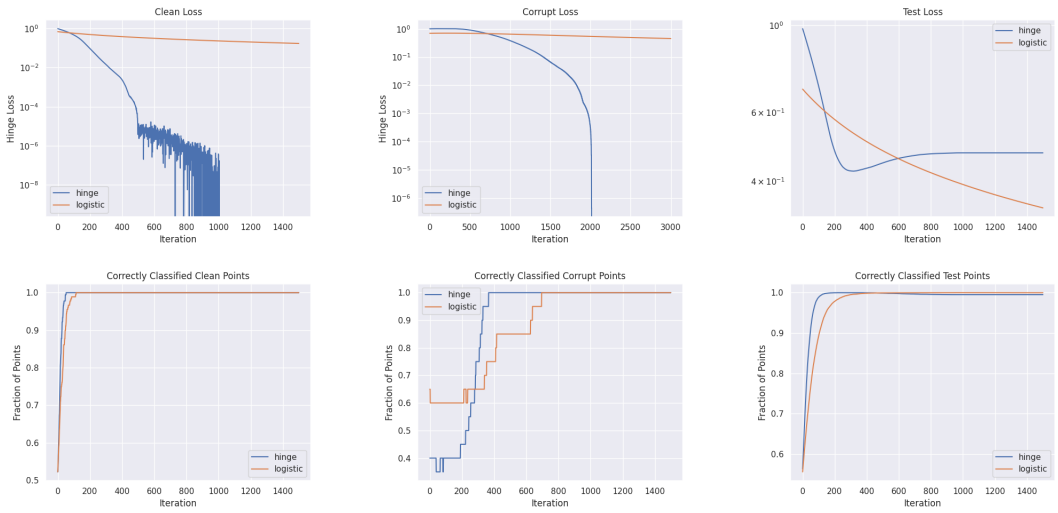

Figure 1: From left to right, the first row shows the clean, corrupt, and test losses as a function of epoch. The second row shows the fraction of clean, corrupt, and test points that are classified correctly. These plots were generated for $n = 100$, $d = 800$, $k/n = 0.1$, $m = 100$, $\gamma = 0.015$, and using gradient descent with a step size of 0.01.

In Figure 1 we call attention to the difference in the training dynamics of hinge loss versus logistic loss. Perhaps the key difference between hinge loss and logistic loss is that the contributions from any given point do not get smaller as the point approaches 0 loss. Furthermore, unlike with the logistic loss points can actually attain zero hinge loss after a finite number of epochs. Once they do attain zero loss, they cease to contribute to the update of the network parameters. If the remaining active points push the parameters in such a way as to increase the loss on a given point then it will reactivate. As a result points close to zero hinge loss periodically activate and deactivate giving rise to the chaotic

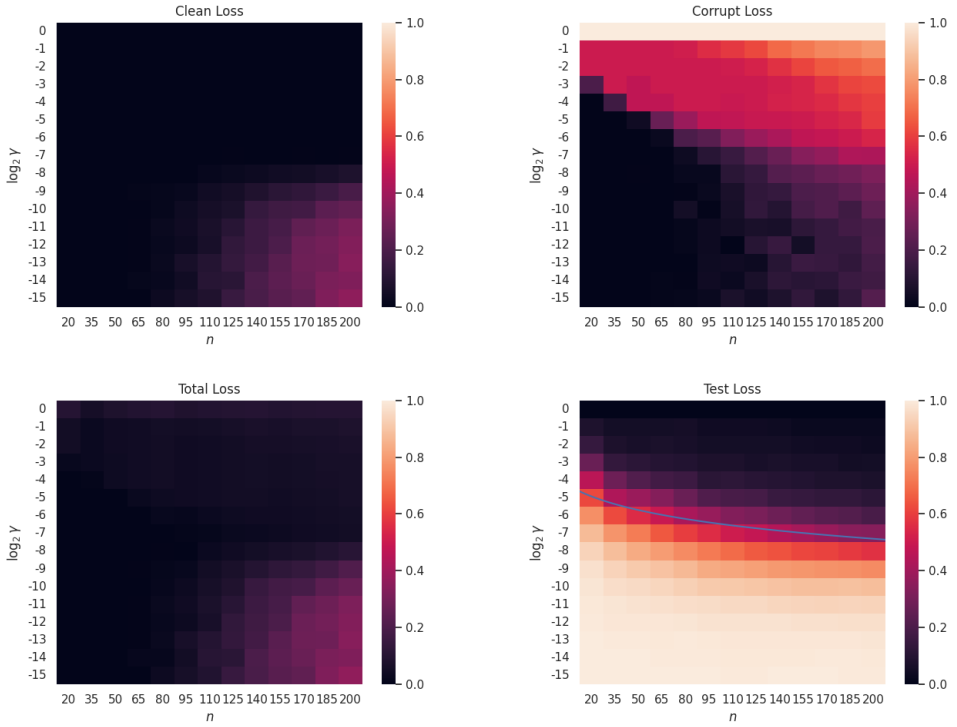

Figure 2: In the top row we show the loss on the clean training points and the corrupt training points after training. In the bottom row we show the total loss after training, along with the test loss on 10000 random generated points after training. For each plot we set $d = 1000, m = 30, \eta = 0.005$ and train for 5000 iterations of gradient descent using hinge loss. In each plot we vary γ and n and hold the fraction of corrupt points constant at 0.05. In the bottom right plot we also graph c/n for $c \approx 0.6$

behavior observed as the training loss approaches zero. We note that managing this behavior required a careful analysis that is distinct from the analysis for logistic loss.

In Figure 2 we call particular attention to the bottom right plot. Our theory predicts a phase transition between benign overfitting and non-benign overfitting when $\gamma \approx c/n$: the phase transition we observe empirically in the bottom-right heatmap suggests this estimate is reasonable. With regard to the hinge loss over the corrupt points, displayed in the top-right heatmap, we observe another phase transition, this time between overfitting and non-overfitting. The top and bottom heatmaps of the left-hand column display the hinge loss over the clean training set and total training set respectively, these appear very similar due to the fact that clean points make up 95% of the training set. The clean points fail to achieve zero, or close to zero, hinge loss only when γ is small and n is large. As stated in the caption, in these experiments d is fixed and thus as n increases the near-orthogonality condition we require on the noise components in order to prove convergence to zero clean loss is compromised. As a result, when γ is small and the correlations between noise vectors is potentially large it is possible for pairs of points with opposite labels to be significantly correlated.

References

Keith Ball. An elementary introduction to modern convex geometry. 1997.

Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. 2015.

Spencer Frei, Niladri S Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pp. 2668–2703. PMLR, 2022.

Xingyu Xu and Yuantao Gu. Benign overfitting of non-smooth neural networks beyond lazy training. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11094–11117. PMLR, 25–27 Apr 2023. URL `https://proceedings.mlr.press/v206/xu23k.html`.