Vision Transformers learn patch association

Anonymous Author(s)
Affiliation
Address
email

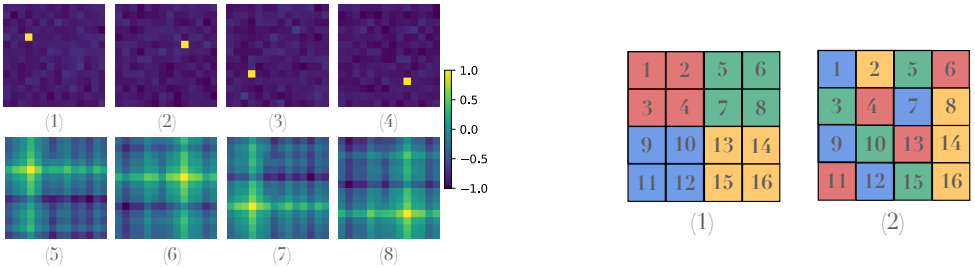
Abstract

Vision Transformers (ViTs) have recently achieved comparable or superior performance to Convolutional neural networks (CNNs) in computer vision. This empirical breakthrough is even more remarkable since ViTs discards spatial information by mixing patch embeddings and positional encodings and do not embed any visual inductive bias (e.g. spatial locality). Yet, recent work showed that while minimizing their training loss, ViTs specifically learn spatially localized patterns. This raises a central question: how do ViTs learn this pattern by solely minimizing their training loss using gradient-based methods from *random initialization*? We propose a structured classification dataset and a simplified ViT model to provide preliminary theoretical justification of this phenomenon. Our model relies on a simplified attention mechanism –the positional attention mechanism– where the attention matrix solely depends on the positional encodings. While the problem admits multiple solutions that generalize, we show that our model implicitly learns the spatial structure of the dataset while generalizing. We finally prove that learning the structure helps to sample-efficiently transfer to downstream datasets that share the same structure as the pre-training one but with different features. We empirically verify that ViTs using only the positional attention mechanism perform similarly to the original one on CIFAR-10/100, SVHN and ImageNet.

1 Introduction

Transformers are deep learning models built on self-attention [38], and in the past several years they have increasingly formed the backbone for state-of-the-art models in domains ranging from Natural Language Processing (NLP) [38, 14] to computer vision [15], reinforcement learning [8, 24], program synthesis [2] and symbolic tasks [27]. Beyond their remarkable performance, several works reported the ability of Transformers to simultaneously minimize their training loss and learn inductive biases tailored to specific datasets e.g. in computer vision [32], in NLP [6, 40] or in mathematical reasoning [43]. In this paper, we focus on computer vision where convolutions are considered to be an adequate and biologically plausible inductive bias since they capture local spatial information [18]. This seems intuitively reasonable: nearby pixels encode the presence of small scale features, whose patterns in turn determine more abstract features at longer and longer length scales. Several seminal works [9, 15, 32] *empirically* show that although randomly initialized, the positional encodings in Vision Transformers (ViTs) [15] learn structured patterns where closer patches have more similar positional encodings – see Figure 1a-(5,6,7,8). We refer to this structure as spatially localized patterns. A priori, learning such specific structure is surprising. Indeed, in contrast to convolutional neural networks (CNNs), ViTs are not built with the inductive bias of spatial locality and weight sharing. They start by replacing an image by a collection of D patches $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_D) \in \mathbb{R}^{d\times D}$, each of dimension d. While each $\boldsymbol{X}_i$ represents (an embedding of) a spatially localized portion of the original image, the relative positions of the patches $\boldsymbol{X}_i$ in the image are disregarded. Instead, relative spatial information is supplied through *image-independent* positional encodings $\boldsymbol{P} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_D) \in \mathbb{R}^{d\times D}$. Unlike CNNs, each layer of a ViT then learns, via trainable self-attention, a non-local set of filters that

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.



(a) (b)

Figure 1: (a) Visualization of the positional encodings similarities at initialization (1-2-3-4) and after training on Imagenet (5-6-7-8) using a "ViT-large-patch16-224" [15]. In contrast with the initial arrays that are random, the final ones show a spatially localized pattern around the patch in yellow. (b) Partition of the patches into sets $\mathcal{S}_\ell$ as in Definition 2.1. Each square depicts a patch along with its index e.g. "1" denotes patch $\boldsymbol{X}_1$. Squares in the same color belong to the same set $\mathcal{S}_\ell$. (1) corresponds to convolutional sets $\mathcal{S}_\ell = \{4(\ell-1)+1, 4(\ell-1)+2, 4(\ell-1)+3, 4\ell\}$ for $\ell \in \{1,2,3,4\}$. Definition 2.1 also covers sets with non-contiguous elements as (2).

non-linearly depend on both the values of all patches $\boldsymbol{X}_j$ and their positional encodings $\boldsymbol{p}_j$. Despite their spatial non-locality, ViTs surprisingly learn the convolution inductive bias in vision datasets as shown in Figure 1a.

This empirical observation sets a central question: from a *theoretical perspective*, how do ViTs manage to learn these spatially localized patterns by simply minimizing their training loss using gradient descent from random initialization? While it is known that attention can express local operations as convolution [9], it remains unclear how ViTs learn it. In this paper, we present a simple spatially-structured classification dataset for which it is sufficient (but not necessary) to learn the structure in order to generalize. We also present a simplified ViT model which we prove implicitly learns a spatially localized inductive bias when it minimizes its training loss via gradient descent (GD). We name this inductive bias *patch association*. We prove that our ViT model leverages this bias to generalize and sample-efficiently transfer to downstream datasets that share the same structure. To our knowledge, our work is the first showing that there exists simple (yet natural) data distributions where simple ViTs implicitly learn spatially localized patterns during training. Our contributions are divided as follows:

- In Section 2, we introduce our framework to formalize patch association.
- In Section 3, we introduce our structured classification dataset and a simplified ViT model to theoretically study how ViTs learn their inductive bias and generalize.
- In Section 4, we present our main theorems (Theorem 4.1 and Theorem 4.2) along with the intermediate lemmas. We prove that a one-layer single-head ViT model trained with GD on our synthetic dataset learns its structure and thus generalizes.
- In Section 5, we show (Theorem 5.1) that after pre-training in our synthetic dataset, our model can be sample-efficiently fine-tuned to transfer to a downstream dataset that shares the same structure as the source dataset (and may have different features).
- In Section 6, we conduct a set of experiments to validate our theoretical model.

Related work

CNNs and ViTs. Many computer vision architectures can be considered as a form of hybridization of Transformers with CNNs. For example, DeTR [7] uses a CNN to generate features that are fed to a Transformer. [16] show that self-attention can be initialized or regularized to behave like a convolution and [11, 20] add convolution operations to Transformers. Conversely, [5, 33, 4] introduce self-attention or attention-like operations to supplement or replace convolution in ResNet-like models. In constrast, our paper attempts to explain how ViTs learn convolution-like patterns using GD.

Theory for attention models. Early theoretical works have focused on the expressivity of attention. [39, 17] addressed this question in the context of self-attention blocks and [12, 41, 22] for Transformers. On the optimization side, [46] investigate the role of adaptive methods in attention models and [35] analyze the dynamics of a single-head attention head to approximate the learning of a Seq2Seq architecture. Our work rather belongs to the implicit bias line of research: we exhibit a setting where a single-head ViT trained with gradient descent provably learns convolution-like patterns and generalizes. To our knowledge, this work is the first that proves generalization of attention models through the lens of optimization.

Notation. We use lower case letters for scalars, lower case bold for vectors and upper case bold for matrices. Given an integer D, we define $[D] = \{1, \dots, D\}$. Any statement made "with high probability" holds with probability at least $1 - 1/\mathrm{poly}(d)$. Given a vector $\boldsymbol{a} \in \mathbb{R}^d$ and $k \leqslant d$, we define $\mathrm{Top}_k\{a_j\}_{j=1}^d = \{a_{i_1}, \dots, a_{i_k}\}$ where $a_{i_1}, \dots, a_{i_k}$ are the k-largest elements. For a function F that implicitly depend on parameters $\boldsymbol{A}$ and $\boldsymbol{v}$, we often write $F_{\boldsymbol{A},\boldsymbol{v}}$ to highlight its parameters.

2 Defining patch association

The goal of this section is to define the notion of patch association. Before, we define the data structure we use to model spatially localized patterns.

Definition 2.1 (Data structure)**.** *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^{d\times D} \times \{-1, 1\}$ where each patch $\boldsymbol{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_D) \in \mathbb{R}^{d\times D}$ has label $y \in \{-1, 1\}$. We partition $[D]$ into L disjoint subsets i.e. $[D] = \bigcup_{\ell=1}^L \mathcal{S}_\ell$ such that $\mathcal{S}_\ell \subsetneq D$ and $|\mathcal{S}_\ell| = C$.*

Note that Definition 2.1 is not an assumption since for any matrix $\boldsymbol{X}$, it is possible to define a partition on its columns. We thus consider distributions $\mathcal{D}$ where there exists an underlying partition of the patches into disjoint sets $\mathcal{S}_\ell$. A particular case of sets $\mathcal{S}_\ell$ is the one of localized sets which we use to model the patterns from Figure 1a. We define them as follows.

Definition 2.2 (Informal: spatially localized sets)**.** *For $\boldsymbol{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_D)$, we say that the sets $\mathcal{S}_\ell$ are localized sets if $\forall \ell \in [L]$, $\forall i \in \mathcal{S}_\ell$, the $\boldsymbol{X}_i$'s form a rectangle.*

An example of spatially localized sets is $\mathcal{S}_1 = \{1, \dots, C\}, \dots, \mathcal{S}_{\lfloor D/C \rfloor} = \{D - C, \dots, D\}$ that we depict in Figure 1b-(1) in the case $D = 16$ and $C = 4$. We emphasize that Definition 2.1 is not limited to spatially localized sets and also covers non-contiguous sets as Figure 1b-(2). Since we aim to show that ViTs learn the sets $\mathcal{S}_\ell$ while fitting the labels y, we need to assume the existence of a labeling function f^* that respects the partition in Definition 2.1.

Assumption 1 (Data structure)**.** *Let $\mathcal{D}$ as in Definition 2.1. We assume that there exists a labeling function f^* such that $\mathbb{P}[yf^*(\boldsymbol{X}) > 0] = 1 - d^{-\omega(1)}$ and,*

$$f^*(\boldsymbol{X}) := \sum_{\ell\in[L]} \phi\big((\boldsymbol{X}_i)_{i\in\mathcal{S}_\ell}\big), \quad \textit{where } \phi\colon \mathbb{R}^{d\times C} \to \mathbb{R} \textit{ is some function,} \tag{1}$$

Assumption 1 states that there exists a labeling function that preserves the underlying structure by applying the same function ϕ to each $\mathcal{S}_\ell$ as in (1). For instance, when the sets $\mathcal{S}_\ell$'s are spatially localized, f^* can be a one-hidden layer convolutional network. In this paper, we are interested in *patch association* which refers to the ability of an algorithm to identify the sets $\mathcal{S}_\ell$'s. Following [9, 15, 32], we investigate whether ViTs can learn patch association using its positional encodings.

Definition 2.3 (Patch association for ViTs)**.** *Let Assumption 1 holds. Let $\mathcal{M}\colon \mathbb{R}^{d\times D} \to \{-1, 1\}$ be a transformer and $\boldsymbol{P}^{(\mathcal{M})}$ its positional encodings matrix. We say that $\mathcal{M}$ performs patch association if for all $\ell \in [L]$ and $i \in \mathcal{S}_\ell$, we have $\mathrm{Top}_C\ \{\langle \boldsymbol{p}_i^{(\mathcal{M})}, \boldsymbol{p}_j^{(\mathcal{M})}\rangle\}_{j=1}^D = \mathcal{S}_\ell$.*

Definition 2.3 states that patch association is learned when for a given $i \in \mathcal{S}_\ell$, its positional encoding mainly attends those of j such that $i, j \in \mathcal{S}_\ell$. In this way, the transformer groups the X_i according to $\mathcal{S}_\ell$ just like the true labeling function – It is thus natural to ask would ViTs really learn those $\mathcal{S}_\ell$ after training to match the labeling function f^*?

ViTs do not always learn patch association under Assumption 1. Figure 1a shows that ViTs learn spatially localized patterns on ImageNet. This raises a question: while minimizing their loss, do ViTs always learn the structure of $\mathcal{S}_\ell$ for any data distribution satisfying Assumption 1? We give a negative answer through the following synthetic experiment. Consider the case where all the patches $\boldsymbol{X}_j$ are i.i.d. standard Gaussian and f^* being a one-hidden layer convolutional network with cubic activation. As shown in Figure 2, one-layer ViT reaches small test error. However, $\boldsymbol{P}^\top\boldsymbol{P}$ does not match the convolution pattern encoded in f^*. To prove that ViTs learn patch association, we need additional assumptions on $\mathcal{D}$.

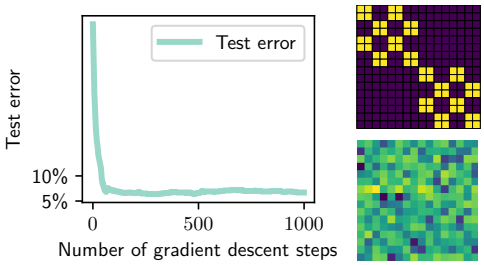


Figure 2: Left: Test error of the ViT on the convolution structure structured dataset. Upper Right: Grid displaying the input patches. Yellow squares represent sets $\mathcal{S}_\ell$ as in Definition 2.2. Those sets are taken into account when computing the convolutional function f^*. Lower Right: Learnt $\boldsymbol{P}^\top\boldsymbol{P}$ looks random compared to upper one.

3 Setting to learn patch association

In this section, we introduce our theoretical setting to analyze how ViTs learn patch association. We first define our binary classification dataset and finally present the ViT model we use to learn it.

Data distribution. The binary classification problem analyzed in this paper is defined as follows.

Definition 3.1 (Data distribution)**.** *Let $\mathcal{D}$ be a distribution satisfying Assumption 1 and $\boldsymbol{w}^* \in \mathbb{R}^d$ be the underlying feature. For each data-point $\boldsymbol{X}$, we define two types of sets:*

- ***Signal set**: uniformly sample an index $\ell(\boldsymbol{X})$ from $[L]$ and for $j \in \mathcal{S}_{\ell(\boldsymbol{X})}$, $\boldsymbol{X}_j = y\boldsymbol{w}^* + \boldsymbol{\xi}_j$, where $y\boldsymbol{w}^*$ is the informative feature and $\boldsymbol{\xi}_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 \mathbf{I}_D)$.*
- ***Random sets**: for $\ell \in [L] \backslash \{\ell(\boldsymbol{X})\}$ and $j \in \mathcal{S}_\ell$, $\boldsymbol{X}_j = \delta_j \boldsymbol{w}^* + \boldsymbol{\xi}_j$, where the spurious feature $\delta_j = 1$ with probability $q/2$, -1 with same probability and 0 otherwise, and $\boldsymbol{\xi}_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 \mathbf{I}_D)$.*

Besides, $\mathcal{D}$ admits the labeling function $f^*(\boldsymbol{X}) = \sum_{\ell \in [L]} \mathrm{Threshold}_{0.9C}(\sum_{i \in \mathcal{S}_\ell} \langle \boldsymbol{w}^*, \boldsymbol{X}_i \rangle)$, where $\mathrm{Threshold}_C(z) = z$ if $|z| > C$ and 0 otherwise.

We sketch a data-point of $\mathcal{D}$ in Figure 3. Our data distribution is a model to describe the patch association phenomenon in Figure 1-(a). Indeed, we can divide each image in sets $\mathcal{S}_\ell$'s that contain adjacent patches. For instance, the patches in yellow in Figure 1-(a) belong to the same $\mathcal{S}_\ell$. Thus, in our setting, proving patch association amounts to show that the positional encodings in the same set $\mathcal{S}_\ell$ are highly correlated. We make the following assumptions on the parameters of $\mathcal{D}$.

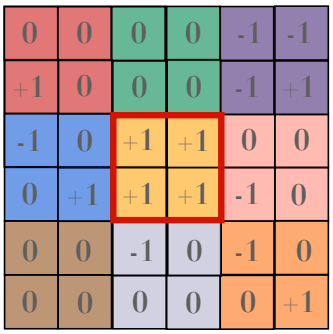


Figure 3: Visualization of a data-point $\boldsymbol{X}$ in $\mathcal{D}$ when the $\mathcal{S}_\ell$'s are spatially localized. Each square depicts a patch $\boldsymbol{X}_j$ and squares of the same color belong to the same set $\mathcal{S}_\ell$. "0" indicates that the patch does not have a feature, "1" stands for feature $1 \cdot \boldsymbol{w}^*$ and "-1" for feature $-1 \cdot \boldsymbol{w}^*$. The large red square depicts the signal set $\ell(\boldsymbol{X})$. Although there are more "-1"'s than "+1"'s, the label of $\boldsymbol{X}$ is $+1$ since there are only "+1"'s inside the signal set.

Assumption 2. *We set $d = \mathrm{poly}(D)$, $C = \mathrm{polylog}(d)$, $q = \mathrm{poly}(C)/D$, $\|\boldsymbol{w}^*\|_2 = 1$ and $\sigma^2 = 1/d$. This implies $C \ll D$ and $q \ll 1$.*

Assumption 2 may be justified by considering a "ViT-base-patch16-224" model [15] on Imagenet. In this case, $d = 384$, $D = 196$. σ is set to have $\|\boldsymbol{\xi}_j\|_2 \approx \|\boldsymbol{w}^*\|_2$. q is chosen so that there are more spurious features than informative ones (low signal-to-noise regime) which makes the data non-linearly separable. Our dataset is thus non-trivial to learn since generalized linear networks fail to generalize.

Theorem 3.1. *Let $\mathcal{D}$ be as defined in Definition 3.1. Let $g(\boldsymbol{X}) = \phi\left(\sum_{j=1}^D \langle \boldsymbol{w}_j, \boldsymbol{X}_j \rangle\right)$ be a generalized linear model. Then, g does not fit the labeling function i.e. $\mathbb{P}[f^*(\boldsymbol{X}) g(\boldsymbol{X}) \leqslant 0] \geqslant 1/8$.*

Intuitively, g fails to generalize because it is does not have any knowledge on the underlying partition and the number of random sets is much higher than the signal one. Thus, a model must have a minimal (but not necessarily complete) knowledge about the $\mathcal{S}_\ell$'s in order to generalize.

Theorem 3.2. *Let $\mathcal{D}$ be defined as in Definition 3.1. There exists a (one-layer) transformer $\mathcal{M}$ so that $\mathbb{P}[f^*(\boldsymbol{X}) \mathcal{M}(\boldsymbol{X}) \leqslant 0] = d^{-\omega(1)}$ but for all $\ell \in [L]$, $i \in \mathcal{S}_\ell$, $\mathrm{Top}_C \left\{\langle \boldsymbol{p}_i^{(\mathcal{M})}, \boldsymbol{p}_j^{(\mathcal{M})} \rangle\right\}_{j=1}^D \cap \mathcal{S}_\ell = \varnothing$.*

Theorem 3.2 states that there exist classifiers that generalize without being able to learn patch association (see Appendix for proof). In Section 4, we show that there exist a ViT model that learns patch association via minimizing its training objective.

Simplified ViT model. We here define our simplified ViT model. Before, we remind the definition of attention [3, 38] in the single-head case with average pooling.

Definition 3.2 (Self-attention)**.** *Let $\boldsymbol{X} \in \mathbb{R}^{d \times D}$ be a patch from $\mathcal{D}$ and $\boldsymbol{P} \in \mathbb{R}^{d \times D}$ the positional encoding. The attention mechanism computes*

1. *the sum of patches and positional encodings i.e. $\boldsymbol{\mathcal{X}} = \boldsymbol{X} + \boldsymbol{P} \in \mathbb{R}^{d \times (D+1)}$.*
2. *the attention matrix $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{K}^\top$ where $\boldsymbol{Q} = \boldsymbol{\mathcal{X}}^\top \boldsymbol{W_Q}$, $\boldsymbol{K} = \boldsymbol{\mathcal{X}}^\top \boldsymbol{W_K}$, $\boldsymbol{W_Q}, \boldsymbol{W_K} \in \mathbb{R}^{d \times d}$.*

3. *the score matrix* $\boldsymbol{S} \in \mathbb{R}^{D\times D}$ *with coefficients* $S_{i,j} = \exp(A_{i,j}/\sqrt{d})/\sum_{r=1}^{D}\exp(A_{i,r}/\sqrt{d})$.
4. *the matrix* $\boldsymbol{V} = \boldsymbol{X}^\top \boldsymbol{W_V}$, *where* $\boldsymbol{W_V} \in \mathbb{R}^{d\times d}$.

It finally outputs $\mathrm{SA}((\boldsymbol{X};\boldsymbol{P})) = \boldsymbol{SV} \in \mathbb{R}^{d\times D}$.

Given its highly non-linear nature with respect to the input, directly analyzing SA is difficult. We now introduce the positional attention mechanism that circumvents this issue.

Definition 3.3 (Positional attention)**.** *Let* $\boldsymbol{X} \in \mathbb{R}^{d\times D}$ *be a datum from* $\mathcal{D}$ *and* $\boldsymbol{P} \in \mathbb{R}^{d\times D}$ *the positional encoding. The positional attention mechanism takes as input the pair* $(\boldsymbol{X};\boldsymbol{P})$ *and computes:*

1. *the attention matrix* $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{K}^\top$ *where* $\boldsymbol{Q} = \boldsymbol{P}^\top \boldsymbol{W_Q}$, $\boldsymbol{K} = \boldsymbol{P}^\top \boldsymbol{W_K}$ *and* $\boldsymbol{W_Q}, \boldsymbol{W_K} \in \mathbb{R}^{d\times d}$.
2. *the matrix* $\boldsymbol{V} = \boldsymbol{X}^\top \boldsymbol{W_V}$, *where* $\boldsymbol{W_V} \in \mathbb{R}^{d\times d}$.

Similarly to self-attention, it computes the score matrix and outputs $\mathrm{PA}((\boldsymbol{X};\boldsymbol{P})) = \boldsymbol{SV}$.

In Definition 3.3, one of the main differences with self-attention is having $\boldsymbol{A} = \boldsymbol{P}^\top \boldsymbol{W_Q}\boldsymbol{W_K}\boldsymbol{P}$. Thus, the advantage of positional attention is to *isolate* the positional encoding from the data. Using this framework, we can thus use $\boldsymbol{A}$ to track the dynamics of $\boldsymbol{P}$ in order to prove patch association as in Figure 1a. On the other hand, the matrix $\boldsymbol{V}$ encodes the data-dependent part. In summary, a model that fits the labeling function and learns patch association in our setting must have i) $\boldsymbol{A}$ learns patch association ii) $\boldsymbol{V}$ learns $\boldsymbol{w}^*$ in order to generalize. Lastly, we make the following simplification in the parameters to ease our analysis.

Simplification 3.1. *In the positional attention mechanism, we set* $d = D$, $\boldsymbol{W_K} = \boldsymbol{I}_D$ *and* $\boldsymbol{W_Q} = \boldsymbol{I}_D$ *which implies* $\boldsymbol{A} = \boldsymbol{P}^\top\boldsymbol{P}$. *Besides, we set* $\boldsymbol{W_V} = [\boldsymbol{v},\dots,\boldsymbol{v}] \in \mathbb{R}^{D\times D}$ *where* $\boldsymbol{v} \in \mathbb{R}^D$.

In Simplification 3.1, we set $\boldsymbol{W_K}$ and $\boldsymbol{W_Q}$ so that $\boldsymbol{A} = \boldsymbol{P}^\top\boldsymbol{P}$. This Gram matrix encodes the inductive bias of the ViT as previously seen in Figure 1a. Besides, since fitting the labeling function requires to learn one feature $\boldsymbol{w}^*$, it is sufficient to parametrize $\boldsymbol{W_V}$ with a vector $\boldsymbol{v}$. Under Simplification 3.1, our simplified ViT model is thus one-hidden attention layer with a single head:

$$F(\boldsymbol{X}) = \sum_{i=1}^{D} \sigma\Big(D\sum_{j=1}^{D} S_{i,j}\langle \boldsymbol{v}, \boldsymbol{X}_j\rangle\Big), \tag{T}$$

where σ is an activation function. Since we aim to the simplest ViT model, we opt for a polynomial activation i.e. $\sigma(x) = x^p + \nu x$ where $p \geqslant 3$ is an odd integer and $\nu = 1/\mathrm{poly}(d)$. Note that this choice of polynomial activation is common in the deep learning theory literature – see e.g. [28, 1, 42] among others. The degree p is odd to make the ViT model compatible with the labeling function and strictly larger than 1 because the data is not linearly separable (Assumption 2). We add a linear part in the activation function to ensure that the gradient is non-zero when $\boldsymbol{v}$ has tiny coefficients. With these simplifications, we formally prove that F is able to learn patch association and generalize.

What aspects of self-attention does positional attention capture? As SA, PA is also permutation-invariant and processes all tokens simultaneously. Besides, PA also computes a score matrix between the different tokens. This similarity matrix is also normalized in a sparse manner with the Softmax operator. The only aspect that PA misses from SA is the fact that $\boldsymbol{S}$ does not depend on the input. Nevertheless, we believe that our PA model encapsulates most of the needed inductive bias in *computer vision* since in Section 6, we observe that PA numerically competes with self-attention.

Trainable parameters. The trainable parameters of our model (T) are specified as follows.

Assumption 3. *In the model* (T)*, we set* $\boldsymbol{A}$ *and* $\boldsymbol{v}$ *as trainable parameters. Besides, without loss of generality, we train all* $A_{i,j}$ *for* $i \neq j$ *and leave the diagonals of* $\boldsymbol{A}$ *fixed.*

Although $\boldsymbol{A} = \boldsymbol{P}^\top\boldsymbol{P}$ and $\boldsymbol{P}$ is trainable, we choose for simplicity to only optimize over $\boldsymbol{A}$. Besides, we leave the $A_{i,i}$'s fixed because the softmax is invariant under the uniform shift of the input.

Learning problems. Given a dataset $\mathcal{Z} = \{(\boldsymbol{X}[i], y[i])\}_{i=1}^N$ sampled from $\mathcal{D}$, we solve the empirical risk minimization problem for the logistic loss defined by:

$$\min_{\widehat{\boldsymbol{A}},\widehat{\boldsymbol{v}}} \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + e^{-y[i]F(\boldsymbol{X}[i])}\right) := \widehat{\mathcal{L}}(\widehat{\boldsymbol{A}},\widehat{\boldsymbol{v}}). \tag{E}$$

Instead of directly analyzing (E), we introduce a proxy where we minimize the population risk

$$\min_{\boldsymbol{A},\boldsymbol{v}} \ \mathbb{E}_{\mathcal{D}}\big[\log\big(1+e^{-yF(\boldsymbol{X})}\big)\big] := \mathcal{L}(\boldsymbol{A},\boldsymbol{v}). \tag{P}$$

We refer to (E) as the *realistic* problem while (P) as the *idealized* problem.

Algorithm. We solve (P) and (E) using gradient descent (GD) for T iterations. The update rule in the case of (P) for $t \in [T]$ and $i, j \in [D]$ is

$$A_{i,j}^{(t+1)} = A_{i,j}^{(t)} - \eta\partial_{A_{i,j}}\mathcal{L}(\boldsymbol{A}^{(t)},\boldsymbol{v}^{(t)}), \quad \boldsymbol{v}^{(t+1)} = \boldsymbol{v}^{(t)} - \eta\nabla_{\boldsymbol{v}}\mathcal{L}(\boldsymbol{A}^{(t)},\boldsymbol{v}^{(t)}), \tag{GD}$$

where $\eta > 0$ is the learning rate. A similar update may be written for (E). We now detail how to set the parameters in (GD).

Parametrization 3.1. *When running GD on (P) and (E), the number of iterations is any $T \geqslant \mathrm{poly}(d)/\eta$. We set the learning rate as $\eta \in \big(0, \frac{1}{\mathrm{poly}(d)}\big)$. The diagonal coefficient of the attention matrix are set for $i \in [D]$ as $A_{i,i}^{(0)} = \widehat{A}_{i,i}^{(0)} = \sigma_{\boldsymbol{A}}\mathbf{I}_D$ where $\sigma_{\boldsymbol{A}} = \mathrm{polyloglog}(d)$. The off-diagonal coefficients of $\boldsymbol{A}$ and the value vector are initialized as:*

1. *Idealized case: $\boldsymbol{v}^{(0)} = \alpha^{(0)}\boldsymbol{w}^*$ where $\alpha^{(0)} = \nu^{1/(p-1)}$ and $A_{i,j}^{(0)} = 0$ for $i \neq j$.*
2. *Realistic case: $\widehat{\boldsymbol{v}}^{(0)} \sim \mathcal{N}(0, \omega^2\mathbf{I}_d)$ and $\widehat{A}_{i,j}^{(0)} \sim \mathcal{N}(0, \omega^2)$ where $i \neq j$ and $\omega = 1/\mathrm{poly}(d)$.*

We now explain our initialization choice in the idealized case. We remind that in Simplification 3.1, we have $\boldsymbol{A} = \boldsymbol{P}^\top\boldsymbol{P}$. If one initializes $\boldsymbol{P} \sim \mathcal{N}(0, \sigma_A\mathbf{I}_D/D)$, then with high probability, $A_{i,i}^{(0)} = \|\boldsymbol{p}_i^{(0)}\|_2^2 = \Theta(\sigma_A)$ and $A_{i,j}^{(0)} = \langle \boldsymbol{p}_i^{(0)}, \boldsymbol{p}_j^{(0)}\rangle = \Theta(\sigma_A/\sqrt{D})$ for $i \neq j$. Since $D \gg 1$, it is then reasonable to set $A_{i,j}^{(0)} = 0$. Besides, we initialize $\boldsymbol{v}^{(0)}$ in $\mathrm{span}(\boldsymbol{w}^*)$ even though this latter should be unknown to the algorithm. We remind that the idealized case is only an easy-to-study proxy to ultimately characterize the realistic dynamics.

4 Learning patch association via matching the labeling function

As announced above, we show that our ViTs (T) implicitly learns patch association and fits the labeling function by minimizing the training objective. We first study the dynamics in (P). Using the analysis in the idealized case, we then characterize the solution found in the realistic problem (E).

4.1 Learning process in the idealized case

In this section, we analyze dynamics of (P). Our main result is that the final iterates of (P) learn patch association while generalize using GD on infinitely many examples sampled from $\mathcal{D}$.

Theorem 4.1. *Assume that we run GD on (P) for T iterations with parameters set as in Parametrization 3.1. With high probability, the ViT model (T)*

1. *learns patch association i.e. for all $\ell \in [L]$ and $i \in \mathcal{S}_\ell$, $\mathrm{Top}_C\ \{A_{i,j}^{(T)}\}_{j=1}^D = \mathcal{S}_\ell$.*
2. *learns the labeling function f^* i.e. $\mathbb{P}_{\mathcal{D}}[f^*(\boldsymbol{X})F_{\boldsymbol{A}^{(T)},\boldsymbol{v}^{(T)}}(\boldsymbol{X}) > 0] \geqslant 1 - o(1)$.*

We now sketch the main ideas to prove the theorem and refer to Appendix D for a complete proof.

Invariance and symmetries. The main feature making (P) appealing is the fact that we take the expectation over $\mathcal{D}$. Indeed, since (T) is permutation-invariant and the data distribution is symmetric, we can thus dramatically simplify the variables in (P). An illustration of this is the next lemma that shows that $\boldsymbol{A}$ can be reduced to three variables in (P).

Lemma 4.1. *There exist $\beta = \sigma_{\boldsymbol{A}}$, $\gamma^{(t)}, \rho^{(t)} \in \mathbb{R}$ such that for all $t \geqslant 0$:*

1. *for all $i \in [D]$, $A_{i,i}^{(t)} = \beta$.*
2. *for all $i, j \in [D]$ such that $i, j \in \mathcal{S}_\ell$ for some $\ell \in [L]$, $A_{i,j}^{(t)} = \gamma^{(t)}$.*
3. *for all $i, j \in [D]$ such that $i \in \mathcal{S}_\ell$ and $j \in \mathcal{S}_m$ for some $\ell, m \in [L]$ with $\ell \neq m$, $A_{i,j}^{(t)} = \rho^{(t)}$.*

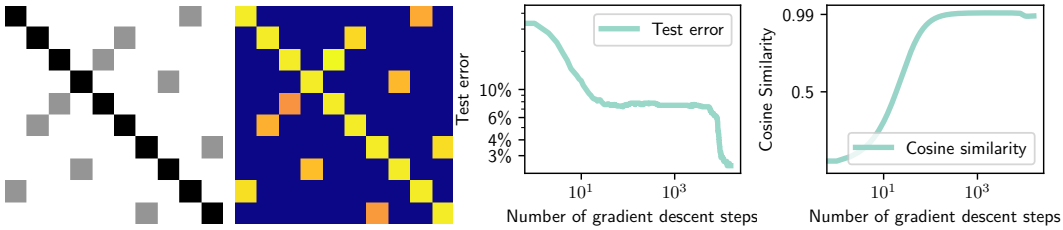


Figure 4: Illustration of Theorem 4.2. From left to right, we first show in grey the tuples (i, j) such that $(i, j) \in \mathcal{S}_\ell$. We then plot the learned matrix $\boldsymbol{A}$ and see that coefficients with high value exactly correspond to their grey scale counterpart in the left plot. We also show test error and cosine similarity of $\boldsymbol{w}^*$ and $\boldsymbol{v}$ w.r.t the number of training steps.

Besides, using the initialization in Parametrization 3.1, we can show that $\boldsymbol{v}$ always lies in $\mathrm{span}(\boldsymbol{w}^*)$.

Lemma 4.2. *For all $t \in [T]$, there exists $\alpha^{(t)} \in \mathbb{R}$ such that $\boldsymbol{v}^{(t)} = \alpha^{(t)} \boldsymbol{w}^*$.*

In summary, Lemma 4.1 and Lemma 4.2 imply that instead of optimizing over $\boldsymbol{A}$ and $\boldsymbol{v}$ in (P), we can instead consider the scalar variables $\alpha^{(t)}$, $\gamma^{(t)}$ and $\rho^{(t)}$. The remaining of this section consists in analyzing the dynamics of these three quantities.

Learning patch association. We first analyze the dynamics of $\gamma^{(t)}$ and $\rho^{(t)}$. To this end, we introduce the following softmax terms:

$$\Lambda^{(t)} = \frac{e^{\beta}}{e^{\beta} + (C-1)e^{\gamma^{(t)}} + (D-C)e^{\rho^{(t)}}}, \qquad \Gamma^{(t)} = \frac{e^{\gamma^{(t)}}}{e^{\beta} + (C-1)e^{\gamma^{(t)}} + (D-C)e^{\rho^{(t)}}},$$

$$\Xi^{(t)} = \frac{e^{\rho^{(t)}}}{e^{\beta} + (C-1)e^{\gamma^{(t)}} + (D-C)e^{\rho^{(t)}}}, \qquad G^{(t)} = D(\Lambda^{(t)} + (C-1)\Gamma^{(t)}).$$

Note that $\Lambda^{(t)}$, $\Gamma^{(t)}$ and $\Xi^{(t)}$ respectively correspond to the coefficients on the diagonal, those for which $i, j \in \mathcal{S}_\ell$ for some ℓ and all the other coefficients of the attention matrix $\boldsymbol{S}$. Using these notations, we first derive the GD updates of $\gamma^{(t)}$ and $\rho^{(t)}$.

Lemma 4.3. *Let $t \leqslant T$. The attention weights $\gamma^{(t)}$ and $\rho^{(t)}$ satisfy:*

$$\gamma^{(t+1)} = \gamma^{(t)} + \eta \mathrm{polylog}(d)(\alpha^{(t)})^p \cdot \Gamma^{(t)}(G^{(t)})^{p-1},$$

$$|\rho^{(t+1)}| \leqslant |\rho^{(t)}| + \eta \mathrm{polylog}(d)(\alpha^{(t)})^p \Big(\frac{1}{D} + \frac{1}{D}\Gamma^{(t)}(G^{(t)})^{p-1}\Big).$$

Lemma 4.3 shows that the increment of $\gamma^{(t)}$ is larger than the one of $\rho^{(t)}$. Since $\gamma^{(0)} = \rho^{(0)} = 0$, this implies that $\gamma^{(t)} \geqslant \rho^{(t)}$ for all $t \geqslant 0$. This observation proves the first item of Theorem 4.1. We now show how learning patch association leads to $\boldsymbol{v}$ highly correlated with $\boldsymbol{w}^*$.

How do ViTs learn patch association? The proof of Theorem 4.1 relies on this mechanism.

- **Event I**: At the beginning of the process, the gradient increment for $\boldsymbol{v}$ is larger than the one of $A_{i,j}^{(t)}$. Because the $A_{i,j}^{(t)}$'s stay very small and $\boldsymbol{v}^{(t)}$ is initialized in $\mathrm{span}(\boldsymbol{w}^*)$, v updates in the direction of $\boldsymbol{w}^*$. At this point, the model is nothing else than a generalized linear model that would not generalize because there are much more noisy tokens than signal ones (see Theorem 3.1).
- **Event II**: Once $\langle \boldsymbol{v}^{(t)}, \boldsymbol{w}^* \rangle$ becomes large, the $A_{i,j}^{(t)}$'s start to update. At this point, because the $A_{i,j}$'s are not negligible, $\boldsymbol{v}^{(t)}$ updates in the direction of $\boldsymbol{w}^*$ and of $\boldsymbol{\xi}_j$. However, the gradient increment for γ (which corresponds to $A_{i,j}^{(t)}$'s in the same set $\mathcal{S}_\ell$) is much larger than the one for ρ (Lemma 4.3). Thus, $\gamma^{(t)}$ increases until being large enough so that the gradient of $\mathcal{L}$ with respect to $\boldsymbol{v}$ is in the direction of $\boldsymbol{w}^*$ again.
- **Event III**: Since the gradient of $\mathcal{L}$ is in the direction of $\boldsymbol{w}^*$, $\boldsymbol{v}^{(t)}$ updates again in the direction of $\boldsymbol{w}^*$ until the population risk becomes a $\Omega(1)$ and we fit the labelling function.

Our mechanism highlights a main advantage of attention-based models: their ability to perform patch association i.e. to set high $A_{i,j}$ for i, j in the same set $\mathcal{S}_\ell$. By doing so, we show that they are able to align their gradient in $\mathrm{span}(\boldsymbol{w}^*)$ and fit the labelling function. Another by-product of this ability is to have sample-efficient transfer as highlighted in Section 5.

4.2 From the idealized to the realistic learning process

The real learning process differs from the idealized one in that i) we have a finite number of samples ii) we initialize both $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{v}}$ as Gaussian random variables. Using a polynomial number of samples, we show that (T) is still able to learn patch association and generalize

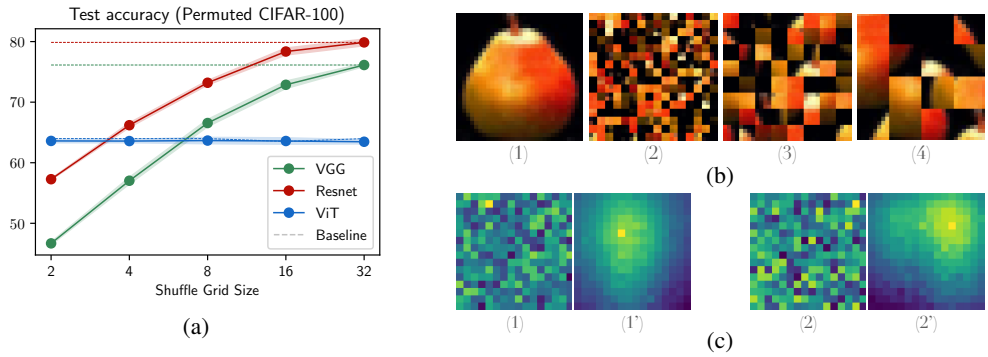


Figure 5: (a): Test accuracy obtained with ViT (patch size 2), ResNet-18 and VGG-19 on permuted (in solid lines) and on original (in dashed lines) CIFAR-100. While convolutional models are very sensitive to permutations, the ViT performs equally whether the dataset is permuted or not. (b): (2) CIFAR-100 image (1) and Permuted CIFAR-100 image when shuffle grid size is 2 (2), 4 (3) and 8 (4). (c): (1-2) Visualization of positional encoding similarities after training a ViT (patch size 2) on permuted CIFAR-100 (shuffle grid size 2). These latter do not have a convolution-like structure. However, when applying the inverse of the permutation, we recover a convolution-like pattern in (1'-2').

Theorem 4.2. *Assume that we run GD on (E) for T iterations with parameters set as in Parametrization 3.1. Assume that the number of samples is $N = \mathrm{poly}(d)$. With high probability, the model*

1. *learns patch association i.e. for all $\ell \in [L]$ and $i \in \mathcal{S}_\ell$, $\mathrm{Top}_C \{\widehat{A}_{i,j}^{(T)}\}_{j=1}^{D} = \mathcal{S}_\ell$.*
2. *fits the labeling function i.e. $\mathbb{P}_{\mathcal{D}}[f^*(\boldsymbol{X}) F_{\widehat{\boldsymbol{A}}^{(T)}, \widehat{\boldsymbol{v}}^{(T)}}(\boldsymbol{X}) > 0] \geqslant 1 - o(1)$.*

Figure 4 numerically illustrates Theorem 4.2. Similarly to [29], the proof introduces a "semi-realistic" learning process that is a mid-point between the idealized and realistic processes. We show that $\widehat{\boldsymbol{A}}^{(T)}$ and $\widehat{\boldsymbol{v}}^{(T)}$ are close to their semi-realistic counterparts – see Appendix E for a complete proof.

5 Patch association yields sample-efficient fine-tuning with ViTs

A fundamental byproduct of our theory is that after pre-training on a dataset sampled from $\mathcal{D}$, our model (T) sample-efficiently transfers to datasets that are structured as $\mathcal{D}$ but differ in their features.

Downstream dataset. Let $\widetilde{\mathcal{D}}$ a downstream data distribution defined as in Definition 3.1 such that its underlying feature is $\widetilde{\boldsymbol{w}}^*$ with $\|\widetilde{\boldsymbol{w}}^*\|_2 = 1$ and $\widetilde{\boldsymbol{w}}^*$ potentially different from $\boldsymbol{w}^*$. In other words, the downstream $\widetilde{\mathcal{D}}$ and source $\mathcal{D}$ distributions share the same structure but not necessarily the same feature. We sample a downstream dataset $\widetilde{\mathcal{Z}} = \{(\widetilde{\boldsymbol{X}}[i], \widetilde{y}[i])\}_{i=1}^{\widetilde{N}}$ from $\widetilde{\mathcal{D}}$.

Learning problem. We consider the model (T) pre-trained as in subsection 4.2. We assume that $\widehat{\boldsymbol{A}}$ is kept *fixed* from the pre-trained model and we only optimize the value vector $\widetilde{\boldsymbol{v}}$ to solve:

$$\min_{\widetilde{\boldsymbol{v}}} \quad \frac{1}{\widetilde{N}} \sum_{i=1}^{\widetilde{N}} \log\left(1 + e^{-\widetilde{y}[i] F(\widetilde{\boldsymbol{X}}[i])}\right) := \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{v}}). \tag{$\widetilde{\text{E}}$}$$

We run GD on ($\widetilde{\text{E}}$) with parameters set as in Parametrization 3.1 except that the $\widehat{A}_{i,j}$'s are fixed and $\widetilde{\boldsymbol{v}}^{(0)} \sim \mathcal{N}(0, \omega^2 \mathbf{I}_d)$ with $\omega = 1/\mathrm{poly}(d)$. Our main results states that this fine-tuning procedure requires a few samples to achieve high test accuracy in $\widetilde{\mathcal{D}}$. In contrast, any algorithm without patch association needs a large number of samples to generalize.

Theorem 5.1. *Let $\widehat{\boldsymbol{A}}$ be the attention matrix obtained after pre-training as in subsection 4.2. Assume that we run GD for T iterations on ($\widetilde{\text{E}}$) to fine-tune the value vector. Using $\widetilde{N} \leqslant \mathrm{polylog}(D)$ samples, the model (T) transfers to $\widetilde{\mathcal{D}}$ i.e. $\mathbb{P}_{\widetilde{\mathcal{D}}}[f^*(\boldsymbol{X}) F_{\widehat{\boldsymbol{A}}, \widetilde{\boldsymbol{v}}^{(T)}}(\boldsymbol{X}) > 0] \geqslant 1 - o(1)$.*

Theorem 5.2. *Let $\mathcal{A}\colon \mathbb{R}^{d \times D} \to \{-1, 1\}$ be a binary classification algorithm without patch association knowledge. Then, it needs $D^{\Omega(1)}$ training samples to get test error $\leqslant o(1)$ on $\widetilde{\mathcal{D}}$.*

The proofs of Theorem 5.1 and Theorem 5.2 are in Appendix F. These theorems hightlight that learning patch association is required for efficient transfer. We believe that they offer a new perspective on explaining why ViTs are widely used in transferring to downstream tasks. While it is possible that ViTs learn shared (with the downstream dataset) features during pretraining, our theory hints that learning the inductive bias of the labeling function is also central for transfer.

6 Numerical experiments

In this section, we empirically verify that i) ViTs learn patch association while miniziming its training loss and ii) the positional attention mechanism competes with the vanilla one on small-scale datasets such as CIFAR-10/100 [26], SVHN [31] and large-scale ones such as ILSVRC-2012 ImageNet [13]. For the small datasets, we use a ViT with 7 layers, 12 heads and hidden/MLP dimension 384. For ImageNet, we train a "ViT-tiny-patch16-224" [15]. Both models are trained with standard augmentations techniques [10] and using AdamW with a cosine learning rate scheduler. We run all the experiments for 300 epochs, with batch size 1024 for Imagenet and 128 otherwise and average our results over 5 seeds. We refer to Appendix A for the training details.

ViTs learn patch association. We consider the CIFAR-100 dataset where we divide each image into grids of size $s \times s$ pixels. For a fixed $s \in \{2, 4, 8, 16, 32\}$, we permute the grids according to π_s to create the permuted CIFAR-100 dataset. We call s the grid shuffle size. Figure 5b-(1) shows a CIFAR-100 image and its corresponding shuffling in the permuted CIFAR-100 dataset Figure 5b-(2-3-4). We train a ViT and convolutional models Resnet18 [21] and VGG-19 [34] on the permuted CIFAR-100 dataset.

For the ViT, we set the patch size to 2, although this is sub-optimal in terms of accuracy, because the patch size needs to stay smaller or equal to s. Figure 5a reports the test accuracy of these three models for different values of s. When s is small, the image does not have a coherent structure e.g. Figure 5b-(2) and thus, convolutional models struggle to generalize. As s increases e.g. Figure 5b-(4), the information inside a patch is meaningful and thus, the convolutional models well-perform. In contrast, ViTs are *insensitive to permutations* as their performance remains unchanged for all s – see Figure 5a. We believe that this is due to the ability of ViTs to learn patch association. Indeed, intuitively, when we permute the grids in Figure 5b, we lose the spatially localized structure and create new sets $\mathcal{S}_\ell$'s and a new labeling function f^*. Despite this change, we verify that the ViT is able to recover the new $\mathcal{S}_\ell$'s: we feed the ViT with the shuffled pear image (Figure 5b-(2)) and see that $\boldsymbol{P}^\top \boldsymbol{P}$ does not have a spatially localized structure in Figure 5c-(1,2). We then apply π_s^{-1} to $\boldsymbol{P}^\top \boldsymbol{P}$ and observe that we recover the spatially localized patterns Figure 5c-(1',2'). This experiment highlights that ViTs do not just group nearby pixels together. They learn patch association even when the $\mathcal{S}_\ell$'s are spatially delocalized.

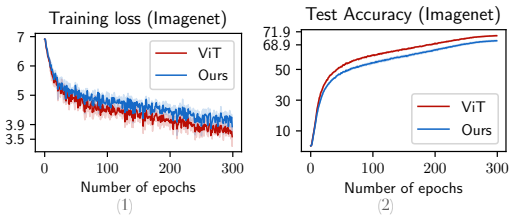


Figure 6: Training loss (1) and test accuracy (2) obtained using a ViT-tiny-patch16-224 on Imagenet. ViT using positional attention (Ours) gets 68.9% test accuracy while vanilla ViT (ViT) gets 71.9%.

ViTs with positional attention are competitive. We numerically verify that ViTs using positional attention compete with those with vanilla attention. In Section 3, we introduced positional attention to define our theoretical learner model. Figure 6 and Figure 7 show that ViTs using positional attention compete with vanilla ViTs on a range of datasets. These experiments strengthen our intuition that for images, having an attention matrix that only depends on the positional encodings is sufficient to have a good test accuracy.

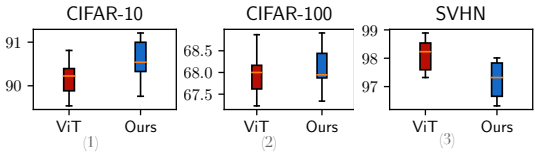


Figure 7: Test accuracy obtained with a ViT using vanilla attention (ViT) and positional attention (Ours) on CIFAR-10 (1), CIFAR-100 (2) and SVHN (3). Our model competes with the vanilla ViT. Patch size 4 and average over 10 seeds for this experiment.

Conclusion, limitations and future works

Our work is a first step towards understanding how Transformers learn tailored inductive biases when trained on large datasets with gradient descent. Our analysis heavily relies on the positional attention mechanism that disentangles patches and positional encodings. In practice, self-attention mixes these two quantities. An interesting direction is to understand the impact of patch embeddings on the inductive bias learned by ViTs. Moreover, our experiment on the Gaussian data shows that ViTs do not always learn the correct inductive bias under Assumption 1: characterizing the distributions under which ViTs recover the structure of the function is an important question. Lastly, this work also paves the way to many extensions beyond convolution. For example, can ViTs learn other inductive bias? What are the inductive biases learnt by Transformers in NLP? Answering those questions is central to better understand the underlying mechanism of attention.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

[2] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] Irwan Bello. Lambdanetworks: Modeling long-range interactions without attention. *arXiv preprint arXiv:2102.08602*, 2021.

[5] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3286–3295, 2019.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[8] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34, 2021.

[9] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2019.

[10] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[11] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021.

[12] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021.

[17] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. *arXiv preprint arXiv:2110.10090*, 2021.

[18] Kunihiko Fukushima. Neocognitron for handwritten digit recognition. *Neurocomputing*, 51:161–180, 2003.

[19] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[20] Jianyuan Guo, Kai Han, Han Wu, Chang Xu, Yehui Tang, Chunjing Xu, and Yunhe Wang. Cmt: Convolutional neural networks meet vision transformers. *arXiv preprint arXiv:2107.06263*, 2021.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[22] Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: Nngp and ntk for deep attention networks. In *International Conference on Machine Learning*, pages 4376–4386. PMLR, 2020.

[23] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016.

[24] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34, 2021.

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[27] Guillaume Lample and François Charton. Deep learning for symbolic mathematics. *arXiv preprint arXiv:1912.01412*, 2019.

[28] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

[29] Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In *Conference on learning theory*, pages 2613–2682. PMLR, 2020.

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[31] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[32] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 2021.

[33] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, 32, 2019.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[35] Charlie Snell, Ruiqi Zhong, Dan Klein, and Jacob Steinhardt. Approximating how single head attention learns. *arXiv preprint arXiv:2103.07601*, 2021.

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[37] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[39] James Vuckovic, Aristide Baratin, and Remi Tachet des Combes. A mathematical theory of attention. *arXiv preprint arXiv:2007.02876*, 2020.

[40] Alex Warstadt and Samuel R Bowman. Can neural networks acquire a structural bias from raw linguistic data? *arXiv preprint arXiv:2007.06761*, 2020.

[41] Colin Wei, Yining Chen, and Tengyu Ma. Statistically meaningful approximation: a case study on approximating turing machines with transformers. *arXiv preprint arXiv:2107.13163*, 2021.

[42] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

[43] Yuhuai Wu, Markus N Rabe, Wenda Li, Jimmy Ba, Roger B Grosse, and Christian Szegedy. Lime: Learning inductive bias for primitives of mathematical reasoning. In *International Conference on Machine Learning*, pages 11251–11262. PMLR, 2021.

[44] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

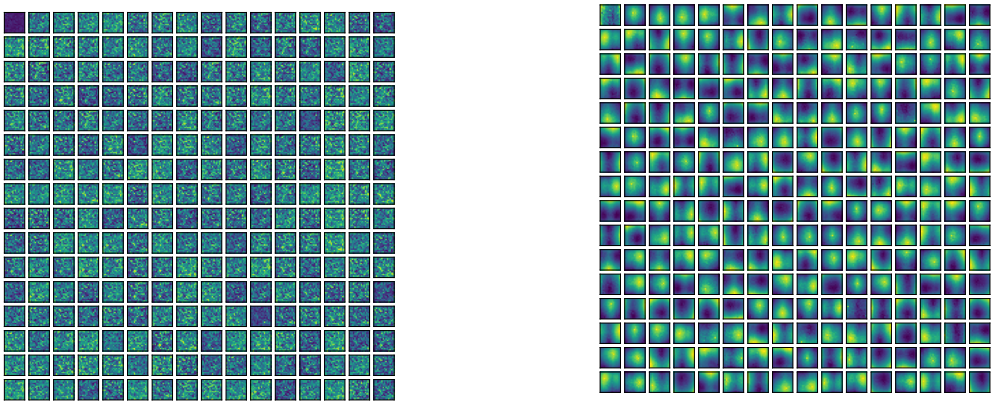
[45] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[46] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

[47] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 4 and Section 5.
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is a theory paper.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 6
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]



(a) (b)

Figure 8: (a) Visualization of the positional encodings similarities when feeding the ViT with Permuted CIFAR-100 data. Each array represents $\{\langle \boldsymbol{p}_i, \boldsymbol{p}_j \rangle\}_{j=1}^{D}$ for a fixed i. (b) displays the positional encodings similarities obtained after inverting the permutation. We see that the ViT is able to recover the convolution-like structure in all the cases.

A Additional experimental details

In this section, we provide additional details on our experiments and additional plots.

A.1 Details on the implementation

We used Pytorch and Nvidia Tesla V100 GPUs. We conduct experiments on small-scale (CIFAR-10/100 and SVHN) and large-scale datasets (ImageNet). The choice of architecture and training parameters depend on the size of the dataset as we detail below.

Small-scale experiments. We use the code available at https://github.com/omihub777/ViT-CIFAR. The model is made of 7 layers, 12 heads, hidden and MLP dimension 384, dropout 0. We use "mean-pooling" and not the CLS pooling. We set the patch size to 2 in the experiment Figure 5 and to 4 in the experiment Figure 7. Indeed, we empirically found that setting patch size 4 was the optimal choice. We apply label smoothing [36] with coefficient 0.1 and do not apply any cutmix [45] nor mixup [44]. We use Adam [25] as optimizer and set the learning rate to 10^{-3}, minimum learning rate to 10^{-5}, β_1 to 0.9, β_2 to 0.999, batch size to 128, weight decay to $5 \cdot 10^{-5}$, number of warmup epochs to 5 and number of total epochs to 200. The scheduler is a cosine learning rate. We used the AutoAugment procedure [10] as in the repository to generate data augmentations. The model has been trained over a single GPU.

Regarding the convolutional models in the experiment Figure 5, we trained a ResNet-18 and a VGG-19 with batch normalization. We trained the two architectures using the same training procedure and hyperparameters as for the ViT.

Large-scale experiments. We use the code available at https://github.com/facebookresearch/deit. Due to limited computational resources, we train a ViT-tiny-patch16-224 [15] where "CLS-pooling" is applied. A detailed table with the hyperparameters used for the ImageNet experiment may be found in Table 9 (column "DeiT-B") in [37]. We set no dropout but set stochastic depth [23] 0.1. We used label smoothing 0.1. Regarding the augmentations, we set RandAugment [10] 9/0.5, mixup 0.8, cutmix 1, erasing probability [47] 0.25. Lastly, we trained the model using AdamW [30] and set the batch size to 1024, learning rate to $5 \cdot 10^{-4} \cdot \frac{\text{batchsize}}{512}$ as in [19], weight decay to 0.05, warmup epochs 5 and number of total epochs to 300. The total number of epochs is 300. The model has been trained over 16 GPUs (8 nodes and 2 GPUs per node) and batch size for each device is 64.

A.2 Additional plots

In Figure 5, we plot the positional encoding similarities for a few patches. Figure 8 provides these plots for all the patches. One should think of Figure 5 as a Figure displaying just two of the arrays present in Figure 8. We consistently verify that the ViT is always able to recover the convolution-like patterns which shows that it is able to learn the right patch association.

B Induction hypothesis

In this section, we present the induction hypothesis that we use in the analysis of the idealized case. This hypothesis is ultimately proved in subsection D.7.

Induction hypothesis B.1. *During the idealized learning process, the following holds for $t \leqslant T$.*

- *the sofmax denominator is large i.e.* $\Lambda^{(t)} + (C-1)\Gamma^{(t)} + (D-C)\Xi^{(t)} = \Theta(D)$.
- $\Xi^{(t)}$ *is not too small i.e.* $\Xi^{(t)} = \Theta(1/D)$.
- $\Gamma^{(t)}$ *and* $\Lambda^{(t)}$ *are in a good range i.e.*

$$\Lambda^{(t)} = \frac{e^{\mathrm{polyloglog}(d)}}{D}, \quad (C-1)\Gamma^{(t)} \in \left[\frac{\Omega(C)}{D}, \frac{\lambda_0}{D}\right],$$

where $\lambda_0 = \Theta(D^{0.01})$.

C Notations

In this section, we introduce the different notations used in the proofs.

General purpose. We first define notations that are used everywhere in the appendix.

- Sigmoid function: Given $z \in \mathbb{R}$, $\mathfrak{S}(z) = 1/(1+e^{-z})$.
- Softmax function: Given $\boldsymbol{z} = (z_1, \dots, z_D) \in \mathbb{R}^D$, $(\mathrm{softmax}(z_1, \dots, z_D))_m = \frac{e^{z_m}}{\sum_{j=1}^D e^{z_j}}$.
- Loss for a data-point $(\boldsymbol{X}, y)$: $L(\boldsymbol{X}) = \log(1 + e^{-yF(\boldsymbol{X})})$.

Analysis in idealized case. We now provide notations used in the analysis of the idealized case.

- $\kappa(\boldsymbol{X}) = \sum_{\ell \neq \ell(\boldsymbol{X})} \sum_{j \in \mathcal{S}_\ell} \delta_j + C$.
- for $i \in [D]$, $\boldsymbol{O}_i^{(t)} = \sum_{j=1}^D S_{i,j}^{(t)} \boldsymbol{X}_j$.

Analysis in realistic case. We now provide notations used in the analysis of the realistic case.

- Score matrix: $\boldsymbol{S} \in \mathbb{R}^{D \times D}$ with coefficients $\widehat{S}_{i,j} = \exp(\widehat{A}_{i,j}/\sqrt{d}) / \sum_{r=1}^D \exp(\widehat{A}_{i,r}/\sqrt{D})$.
- Given a data-point $(\boldsymbol{X}[i], y[i])$ and $j \in [D]$, $\boldsymbol{O}_j^{(t)}[i] = \sum_{k=1}^D S_{j,k}^{(t)} \boldsymbol{X}_k[i]$

D Learning process in the idealized setting

D.1 Roadmap of the proof

From Lemma 4.2, we know that $\boldsymbol{v}^{(t)} \in \mathrm{span}(\boldsymbol{w}^*)$ for all $t \in [T]$. The main idea of the proof consists in analyzing the GD dynamics of $\alpha^{(t)} = \langle \boldsymbol{v}^{(t)}, \boldsymbol{w}^* \rangle$ that satisfy

$$\alpha^{(t+1)} = \alpha^{(t)} + \eta \underbrace{\mathbb{E}\left[y\mathfrak{S}(-yF_{\boldsymbol{A}^{(t)},\boldsymbol{v}^{(t)}}(\boldsymbol{X})) \sum_{i \in \mathcal{S}_{\ell(\boldsymbol{X})}} \sigma'\Big(\sum_{j=1}^D S_{i,j}^{(t)} \langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_j \rangle \Big) \sum_{k=1}^D S_{i,k}^{(t)} \langle \boldsymbol{w}^*, \boldsymbol{X}_m \rangle \right]}_{\mathcal{S}^{(t)}}$$
$$+ \eta \underbrace{\mathbb{E}\left[y\mathfrak{S}(-yF_{\boldsymbol{A}^{(t)},\boldsymbol{v}^{(t)}}(\boldsymbol{X})) \sum_{i \notin \mathcal{S}_{\ell(\boldsymbol{X})}} \sigma'\Big(\sum_{j=1}^D S_{i,j}^{(t)} \langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_j \rangle \Big) \sum_{m=1}^D S_{i,m}^{(t)} \langle \boldsymbol{w}^*, \boldsymbol{X}_m \rangle \right]}_{\mathcal{N}^{(t)}}. \qquad \text{(GD-}\alpha\text{)}$$

We divide the idealized learning process as follows.

– **Event I** ($t \in [0, \mathcal{T}_0]$, subsection D.2): at initialization, $\alpha^{(0)}$ is small. Therefore, the sigmoid $\mathfrak{S}(-yF(\boldsymbol{X}))$ is large. Besides, around $\alpha^{(0)}$, it stays constant i.e. $\mathfrak{S}(-yF_{\boldsymbol{A}^{(t)},\boldsymbol{v}^{(t)}}(\boldsymbol{X})) \approx \mathfrak{S}(-yF_{\boldsymbol{A}^{(0)},\boldsymbol{v}^{(0)}}(\boldsymbol{X}))$ in (GD-α). This implies that $\boldsymbol{\mathcal{N}}^{(t)} = 0$ which yields $\alpha^{(t)}$ to increase until reaching a specific value where the sigmoid is not constant anymore.

– **Event II** ($t \in [\mathcal{T}_0, \mathcal{T}_1]$, subsection D.3): at time $\mathcal{T}_0$, $\alpha^{(t)}$ is large. This fact along with Lemma 4.3 imply that $\gamma^{(t)}$ increases. Eventually, $\Gamma^{(\mathcal{T}_1)}$ becomes large enough so that $\boldsymbol{\mathcal{S}}^{(t)} \geqslant \max_{\tau \leqslant T} |\boldsymbol{\mathcal{N}}^{(\tau)}|$.

– **Event III** ($t \in [\mathcal{T}_1, T]$, subsection D.4): Since $\boldsymbol{\mathcal{S}}^{(t)} \geqslant \max_{\tau \leqslant T} |\boldsymbol{\mathcal{N}}^{(\tau)}|$, $\alpha^{(t)}$ increases again. It increases until the population risk is at most $o(1)$.

After T iterations, $\alpha^{(t)}$ is large and the population risk thus converges (subsection D.5). Since the logistic loss is a surrogate for the 0-1 loss, we prove that the learner model fits the labeling function (subsection D.6) which implies the first statement of Theorem 4.1.

Remark : Since we initialize $\alpha^{(0)} \geqslant \nu^{1/(p-1)}$, Lemma D.7 implies that we can overlook the linear part of the activation in this section. Therefore, we only consider $\sigma(x) = x^p$ in the idealized process.

D.2 Event I: $\alpha^{(t)}$ initially increases

A first question that arises is: starting from $\alpha^{(0)}$, what is the value of $\alpha^{(t)}$ that makes the sigmoid non-constant? The following lemma addresses this question.

Lemma D.1. *The value $\alpha^{(t)}$ at which the sigmoid $\mathfrak{S}(-yF(\boldsymbol{X}))$ becomes non-constant is:*

$$\tilde{\alpha} = \frac{\Theta(1)}{C^2 \lambda_0}.$$

Proof of Lemma D.1. The update of $\alpha^{(t)}$ is

$$\begin{aligned} \alpha^{(t+1)} &= \alpha^{(t)} + \eta \mathbb{E}[y\mathfrak{S}(-yF(\boldsymbol{X}))\mathscr{G}^{(t)}] \\ &= \alpha^{(t)} + \eta \mathbb{E}[y\mathfrak{S}(-\alpha^{(0)})\mathscr{G}^{(t)}] + \eta \mathbb{E}\big[y\big(\mathfrak{S}(-yF(\boldsymbol{X})) - \mathfrak{S}(-\alpha^{(0)})\big)\mathscr{G}^{(t)}\big], \end{aligned} \tag{2}$$

where $\mathscr{G}^{(t)} := \sum_{j=1}^{D} \langle \boldsymbol{O}_j^{(t)}, \boldsymbol{w}^* \rangle^p$. Since $x \mapsto \mathfrak{S}(-x)$ is $1/4$-Lipschitz, we rewrite (2) as:

$$\begin{aligned} \big|\alpha^{(t+1)} - \alpha^{(t)} - \eta \mathbb{E}[y\mathfrak{S}(-\alpha^{(0)})\mathscr{G}^{(t)}]\big| &\leqslant \frac{\eta}{4} \mathbb{E}\big[|F(\boldsymbol{X}) - y\alpha^{(0)}| \cdot |y\mathscr{G}^{(t)}|\big] \\ &\leqslant \Theta(C\eta)(\alpha^{(t)})^p \mathbb{E}\big[(C^{1+1/p}\lambda_0)^p |y\mathscr{G}^{(t)}|\big] \\ &\leqslant \Theta(\eta)(\alpha^{(t)})^p \mathbb{E}\big[(C^2\lambda_0)^p |y\mathscr{G}^{(t)}|\big], \end{aligned} \tag{3}$$

where we applied Lemma D.3, Induction Hypothesis B.1 and the fact that $\alpha^{(0)}$ is small in the penultimate inequality. Using Lemma D.4 and Lemma D.5, we have $\mathbb{E}[|y\mathscr{G}^{(t)}|] \leqslant \Theta(1)\mathbb{E}[y\mathscr{G}^{(t)}]$ which yields

$$\big|\alpha^{(t+1)} - \alpha^{(t)} - \eta \mathbb{E}[y\mathfrak{S}(-\alpha^{(0)})\mathscr{G}^{(t)}]\big| \leqslant \Theta(\eta)(\alpha^{(t)})^p \mathbb{E}\big[(C^2\lambda_0)^p y\mathscr{G}^{(t)}\big]. \tag{4}$$

(3) shows that when $\alpha^{(t)}$ is small, we have $\alpha^{(t+1)} \approx \alpha^{(t)} + \eta \mathbb{E}[y\mathfrak{S}(-\alpha^{(0)})A]$. Besides, we have $\mathbb{E}[y\mathscr{G}^{(t)}] \geqslant 0$ so $\alpha^{(t)}$ increases. However, during this increase, the right-hand side of (3) increases and this approximation does not hold anymore. Therefore, the sigmoid is approximately constant when $\alpha^{(t)}$ satisfies:

$$\Theta(1)(\alpha^{(t)})^p \mathbb{E}[(C^2\lambda_0)^p y\mathscr{G}^{(t)}] \leqslant \mathbb{E}[\mathfrak{S}(-\alpha^{(0)})y\mathscr{G}^{(t)}] \iff \alpha^{(t)} \leqslant \frac{\Theta(1)}{C^2\lambda_0}. \tag{5}$$

□

Lemma D.2. *Let $\mathcal{T}_0 = \Theta\left(\frac{1}{\eta C(\alpha^{(0)})^{p-1}}\right)$. For all $t \in [0, \mathcal{T}_0]$, we have $\boldsymbol{\mathcal{N}}^{(t)} = 0$. Therefore, $\alpha^{(t)}$ is updated as*

$$\alpha^{(t+1)} = \alpha^{(t)} + \Theta(\eta C)(G^{(t)})^p (\alpha^{(t)})^{p-1}.$$

Consequently, $\alpha^{(t)}$ is non-decreasing and after $\mathcal{T}_0$ iterations, we have $\alpha^{(t)} \geqslant \frac{\Omega(1)}{C^2\lambda_0}$ for $t \geqslant \mathcal{T}_0$.

Proof of Lemma D.2. For $t \in [0, \mathcal{T}_0]$, we know that the sigmoid $\mathfrak{S}(-yF(\boldsymbol{X}))$ is constant. We apply Lemma D.4 and Lemma D.6 to respectively bound $\boldsymbol{\mathcal{S}}^{(t)}$ and $\boldsymbol{\mathcal{N}}^{(t)}$ in the update of $\alpha^{(t)}$.

$$\alpha^{(t+1)} = \alpha^{(t)} + \Theta(\eta C)(\alpha^{(t)})^{p-1}(G^{(t)})^p. \tag{6}$$

We apply Induction Hypothesis B.1 and $\Omega(C) \geqslant 0$ and $e^{\mathrm{polyloglog}(d)} \leqslant \lambda_0$ in (6) and obtain:

$$\begin{cases} \alpha^{(t+1)} \geqslant \alpha^{(t)} + \Theta(\eta C) e^{\mathrm{polyloglog}(d)} (\alpha^{(t)})^{p-1} \\ \alpha^{(t+1)} \leqslant \alpha^{(t)} + \Theta(\eta C) \lambda_0^p (\alpha^{(t)})^{p-1} \end{cases}. \tag{7}$$

(7) indicates that $\alpha^{(t)}$ is a non-decreasing sequence. Therefore, there exists a time $\mathcal{T}_0$ such that $\alpha^{(\mathcal{T}_0)} = \frac{\Theta(1)}{C^2 \lambda_0}$. Using Lemma K.1, the time $\mathcal{T}_0$ is equal to:

$$\mathcal{T}_0 = \frac{1}{\eta C (\alpha^{(0)})^{p-1} e^{\mathrm{polyloglog}(d)}} + (\lambda_0)^p e^{-\mathrm{polyloglog}(d)} \Big[-\log\big(C^2 \lambda_0 \alpha^{(0)}\big) \Big]. \tag{8}$$

□

D.2.1 Auxiliary lemmas

In this section, we present the auxiliary lemmas needed to prove the main results of subsection D.2. We first present a lemma that bounds the learner model.

Lemma D.3. *Let $t \in [0, T]$. The learner model F is bounded for all $(\boldsymbol{X}, y) \sim \mathcal{D}$ as:*

$$yF_{\boldsymbol{A}^{(t)}, \boldsymbol{v}^{(t)}}(\boldsymbol{X}) \leqslant \Theta(1)(C^{1+1/p} \alpha^{(t)} \lambda_0)^p.$$

Proof of Lemma D.3. By definition, the learner model is:

$$yF_{\boldsymbol{A}^{(t)}, \boldsymbol{v}^{(t)}}(\boldsymbol{X}) = y \sum_{i=1}^{D} \langle \boldsymbol{v}^{(t)}, \boldsymbol{O}_i^{(t)} \rangle^p = \sum_{i \in \mathcal{S}_{\ell(X)}} \langle \boldsymbol{v}^{(t)}, \boldsymbol{O}_i^{(t)} \rangle^p + \sum_{i \notin \mathcal{S}_{\ell(X)}} \langle v^{(t)}, \boldsymbol{O}_i^{(t)} \rangle^p. \tag{9}$$

We successively apply Lemma D.4, Induction Hypothesis B.1 and Lemma D.5 to bound (9).

$$yF_{\boldsymbol{A}^{(t)}, \boldsymbol{v}^{(t)}}(\boldsymbol{X}) \leqslant (\alpha^{(t)})^p \Big(C\Theta(DC\Gamma^{(t)})^p + o(\lambda_0)^p \Big). \tag{10}$$

Finally, we apply Induction Hypothesis B.1 in (10) to obtain the desired result. □

We now present lemmas that bound $\boldsymbol{\mathcal{S}}^{(t)}$ and $\boldsymbol{\mathcal{N}}^{(t)}$.

Lemma D.4. *Let $t \in [T]$ and $i \in \mathcal{S}_{\ell(\boldsymbol{X})}$. We have $y\langle \boldsymbol{w}^*, \boldsymbol{O}_i^{(t)} \rangle^p = \Theta(G^{(t)})^p$. As long as the population risk is not $o(1)$, we have $\boldsymbol{\mathcal{S}}^{(t)} = C\Theta(G^{(t)})^p$ for all $t \leqslant T$.*

Proof of Lemma D.4. We apply Lemma G.3 and obtain:

$$\begin{aligned} y\langle \boldsymbol{w}^*, \boldsymbol{O}_i^{(t)} \rangle^p &= D^p \Bigg((\Lambda^{(t)} + (C-1)\Gamma^{(t)}) + y\Xi^{(t)} \sum_{h \in [L] \setminus \{\ell(\boldsymbol{X})\}} \sum_{r \in \mathcal{S}_h} \delta_r \Bigg)^p \\ &= \Theta\Big(D\big(\Lambda^{(t)} + (C-1)\Gamma^{(t)}\big) \Big)^p \\ &= \Theta(G^{(t)})^p. \end{aligned}$$

□

Lemma D.5. *Let $t \in [T]$. We have $\sum_{i \notin \mathcal{S}_{\ell(\boldsymbol{X})}} y\langle \boldsymbol{w}^*, \boldsymbol{O}_i^{(t)} \rangle^p \leqslant o(\lambda_0)^p$. In particular, this implies $\boldsymbol{\mathcal{N}}^{(t)} \leqslant o(\lambda_0)^p$ for all $t \leqslant T$.*

Proof of Lemma D.5. We have:

$$\begin{aligned}&\sum_{k\in[L]\setminus\{\ell(X)\}}\sum_{i\in\mathcal{S}_k} y\langle \boldsymbol{w}^*, \boldsymbol{O}_i^{(t)}\rangle^p\\ =&D^p\sum_{k\in[L]\setminus\{\ell(X)\}}\sum_{m\in\mathcal{S}_k} y\Big(\Lambda^{(t)}\delta_m+\Gamma^{(t)}\sum_{s\in\mathcal{S}_k\setminus\{m\}}\delta_s+\Xi^{(t)}\Big[Cy+\sum_{h\in[L]\setminus\{\ell(X),k\}}\sum_{r\in\mathcal{S}_h}\delta_r\Big]\Big)^p.\end{aligned}\tag{11}$$

We distinguish two cases.

- $\delta_r=0$ for all $r\in\mathcal{S}_\ell$: we apply Lemma K.3 and obtain:

$$\begin{aligned}\sum_{k\neq\ell(\boldsymbol{X})}\sum_{i\in\mathcal{S}_k} y\langle \boldsymbol{w}^*, \boldsymbol{O}_i^{(t)}\rangle^p =&\sum_{k\neq\ell(\boldsymbol{X})}\sum_{m\in\mathcal{S}_k} y\Big(D\Xi^{(t)}\Big[Cy+\sum_{h\neq\{\ell(\boldsymbol{X}),k\}}\sum_{r\in\mathcal{S}_h}\delta_r\Big]\Big)^p\\ &\leqslant D\cdot\Theta(D\Xi^{(t)}qD\log(d))^p.\end{aligned}\tag{12}$$

- $\exists r\in\mathcal{S}_\ell$ such that $\delta_r\neq 0$: let $i\in\mathcal{S}_k$. We apply Lemma K.4 and obtain:

$$y\langle \boldsymbol{w}^*, \boldsymbol{O}_i^{(t)}\rangle^p\leqslant\Theta\big(D(\Lambda^{(t)}+O(1)\Gamma^{(t)})\big)^p\cdot\mathbf{1}_{\exists r\in\mathcal{S}_k,\delta_r\neq 0}.\tag{13}$$

  We now sum (13) and apply Lemma K.3 to obtain:

$$\begin{aligned}\sum_{k\neq\ell(\boldsymbol{X})}\sum_{i\in\mathcal{S}_k} y\langle \boldsymbol{w}^*, \boldsymbol{O}_i^{(t)}\rangle^p\leqslant&\Theta\big(D(\Lambda^{(t)}+O(1)\Gamma^{(t)})\big)^p\sum_{k\neq\ell(\boldsymbol{X})}\sum_{i\in\mathcal{S}_k}|\delta_r|\\ \leqslant&\Theta\big(D(\Lambda^{(t)}+O(1)\Gamma^{(t)})\big)^p qD\log(d).\end{aligned}\tag{14}$$

  We finally apply Induction Hypothesis B.1 to have $\Gamma^{(t)}\leqslant\lambda_0/C$ in (14) and get

$$\sum_{k\neq\ell(\boldsymbol{X})}\sum_{i\in\mathcal{S}_k} y\langle \boldsymbol{w}^*, \boldsymbol{O}_i^{(t)}\rangle^p\leqslant\Theta(\lambda_0/C)^p qD\log(d).\tag{15}$$

We finally plug (12) and (15) in (11) and obtain:

$$\sum_{k\neq\ell(\boldsymbol{X})}\sum_{i\in\mathcal{S}_k} y\langle \boldsymbol{w}^*, \boldsymbol{O}_i^{(t)}\rangle^p\leqslant D\cdot\Theta(D\Xi^{(t)}qD\log(d))^p+\Theta(\lambda_0/C)^p qD\log(d)\leqslant o(\lambda_0)^p.$$

□

Lemma D.6. *Let $t\in[0,\mathcal{T}_0]$. We have $\boldsymbol{\mathcal{N}}^{(t)}=0$.*

Proof of Lemma D.6. By definition of $\boldsymbol{\mathcal{N}}^{(t)}$, we have:

$$\begin{aligned}&\boldsymbol{\mathcal{N}}^{(t)}\\ =&\frac{1}{2}\Big(\mathbb{E}\Big[\mathfrak{S}(-F(\boldsymbol{X}))\sum_{j\notin\mathcal{S}_{\ell(\boldsymbol{X})}}\langle \boldsymbol{w}^*, \boldsymbol{O}_j^{(t)}\rangle^p\Big|y=1\Big]-\mathbb{E}\Big[\mathfrak{S}(F(\boldsymbol{X}))\sum_{j\notin\mathcal{S}_{\ell(\boldsymbol{X})}}\langle \boldsymbol{w}^*, \boldsymbol{O}_j^{(t)}\rangle^p\Big|y=-1\Big]\Big)\\ =&\Theta(1)\Big(\mathbb{E}\Big[\sum_{j\notin\mathcal{S}_{\ell(\boldsymbol{X})}}\langle \boldsymbol{w}^*, \boldsymbol{O}_j^{(t)}\rangle^p\Big|y=1\Big]-\mathbb{E}\Big[\sum_{j\notin\mathcal{S}_{\ell(\boldsymbol{X})}}\langle \boldsymbol{w}^*, \boldsymbol{O}_j^{(t)}\rangle^p\Big|y=-1\Big]\Big),\end{aligned}\tag{16}$$

where we use $\mathfrak{S}(-F(\boldsymbol{X}))\approx\mathfrak{S}(-F_{\boldsymbol{A}^{(0)},\boldsymbol{v}^{(0)}}(\boldsymbol{X}))$ for $t\in[0,\mathcal{T}_0]$ in the last equality of (16). We now show that each of the summands in (16) is zero. Without loss of generality, let's focus on the first summand. The same reasoning holds for the second one. In particular, for $j\in\mathcal{S}_k$ with $k\neq j$, we now compute $\mathbb{E}[\langle \boldsymbol{w}^*, \boldsymbol{O}_j^{(t)}\rangle^p|y=1]$. Using the binomial theorem and the independence of the δ_r's, we have:

$$\begin{aligned}&\mathbb{E}[\langle \boldsymbol{w}^*, \boldsymbol{O}_j^{(t)}\rangle^p|y=1]\\ =&\sum_{a=0}^{p}\binom{p}{a}(\Lambda^{(t)})^{p-a}\mathbb{E}[(\delta_j)^{p-a}]\mathbb{E}\Big[\Big(\Gamma^{(t)}\sum_{s\in\mathcal{S}_k\setminus\{j\}}\delta_s+\Xi^{(t)}\Big[C+\sum_{h\neq\{\ell(\boldsymbol{X}),k\}}\sum_{r\in\mathcal{S}_h}\delta_r\Big]\Big)^a\Big].\end{aligned}\tag{17}$$

For a even, we have $p-a$ odd which implies $\mathbb{E}[(\delta_u)^{p-a}] = 0$. Therefore, the summands with even a are zero. We now focus on the case a odd. We again apply the binomial theorem and the independence of the δ_r's to get:

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\Gamma^{(t)}\sum_{s\in\mathcal{S}_k\setminus\{j\}}\delta_s + \Xi^{(t)}\Big[C + \sum_{h\neq\{\ell(\boldsymbol{X}),k\}}\sum_{r\in\mathcal{S}_h}\delta_r\Big]\Big)^a\Big] \\
&= \sum_{b=0}^{a}\binom{a}{b}(\Gamma^{(t)})^{a-b}\mathbb{E}\Big[\Big(\sum_{s\in\mathcal{S}_k\setminus\{j\}}\delta_s\Big)^{a-b}\Big](\Xi^{(t)})^b\mathbb{E}\Big[\Big(Cy + \sum_{h\neq\{\ell(\boldsymbol{X}),k\}}\sum_{r\in\mathcal{S}_h}\delta_r\Big)^b\Big].
\end{aligned} \tag{18}$$

For b even, we have $a-b$ odd. This implies that $\mathbb{E}\big[\big(\sum_{s\in\mathcal{S}_k\setminus\{j\}}\delta_s\big)^{a-b}\big] = 0$. In the case b odd, we exactly use the same argument and obtain:

$$\mathbb{E}\Big[\Big(Cy + \sum_{h\neq\{\ell(\boldsymbol{X}),k\}}\sum_{r\in\mathcal{S}_h}\delta_r\Big)^b\Big] = 0. \tag{19}$$

(19) implies (18) is zero and which lastly implies (17) is zero. We conclude that $\mathbb{E}[\langle \boldsymbol{w}^*, \boldsymbol{O}_j^{(t)}\rangle^p | y = 1] = \mathbb{E}[\langle \boldsymbol{w}^*, \boldsymbol{O}_j^{(t)}\rangle^p | y = -1] = 0$ and thus $\boldsymbol{\mathcal{N}}^{(t)} = 0$.

□

Lemma D.7. *Let $(\boldsymbol{X}, \cdot) \sim \mathcal{D}$ and $j \in [D]$. Assume that $\alpha^{(t)} \geqslant \nu^{1/(p-1)}$. Then, we have:*

$$\sigma'\Big(D\sum_{k=1}^{D}\boldsymbol{S}_{j,k}^{(t)}\langle \boldsymbol{v}, \boldsymbol{X}_k\rangle\Big) = \Theta(1)\Big(D\sum_{k=1}^{D}\boldsymbol{S}_{j,k}^{(t)}\langle \boldsymbol{v}, \boldsymbol{X}_k\rangle\Big)^{p-1}.$$

Proof of Lemma D.7. We remind that the derivative of the activation function $\sigma'(x) = px^{p-1} + \nu$. We first remark that for all x, $\sigma'(x) \geqslant px^{p-1}$. Besides, we have:

$$px^{p-1} + \nu \leqslant \frac{3px^{p-1}}{2} \iff x \geqslant \Big(\frac{2\nu}{p}\Big)^{1/(p-1)}. \tag{20}$$

In our case, we have $x = D\alpha^{(t)}\sum_{k=1}^{D}\boldsymbol{S}_{j,k}^{(t)}\langle \boldsymbol{w}^*, \boldsymbol{X}_k\rangle$. Using Induction Hypothesis B.1, we have $x \leqslant C\lambda_0\alpha^{(t)}$. Therefore, a sufficient condition for (20) to hold is $\alpha^{(t)} \geqslant \nu^{1/(p-1)}/(C\lambda_0)$. Since $\lambda_0, C \ll \mathrm{poly}(d)$, we can simplify this condition as $\alpha^{(t)} \geqslant \nu^{1/(p-1)}$.

□

D.3 Event II: $\Gamma^{(t)}$ significantly increases

In this section, we show the increase of $\alpha^{(t)}$ for $t \in [0, \mathcal{T}_0]$ leads to the increase of $\Gamma^{(t)}$. At time $\mathcal{T}_1 > \mathcal{T}_0$, $\Gamma^{(t)}$ is significantly large.

Lemma D.8. *Let $\mathcal{T}_1 = \mathcal{T}_0 + \Theta\Big(\frac{\lambda_0^p}{\eta e^{\mathrm{polyloglog}(d)}}\Big)$. For all $t \in [\mathcal{T}_1, T]$, we have $\Gamma^{(t)} \geqslant \frac{\Omega(\lambda_0)}{D}$. This implies $\boldsymbol{\mathcal{S}}^{(t)} \geqslant \max_{\tau\leqslant T}|\boldsymbol{\mathcal{N}}^{(\tau)}|$.*

Proof of Lemma D.8. Let $t \in [\mathcal{T}_0, T]$ and $\tau \in [\mathcal{T}_0, t]$. Using Corollary G.1 and Induction Hypothesis B.1, $\gamma^{(\tau)}$ satisfies:

$$\gamma^{(\tau+1)} \geqslant \gamma^{(\tau)} + \Omega(\eta C)(\alpha^{(\tau)})^p e^{\mathrm{polyloglog}(d)}. \tag{21}$$

Summing (21) for $\tau = \mathcal{T}_0, \ldots, t-1$ yields

$$\gamma^{(t)} \geqslant \gamma^{(\mathcal{T}_0)} + e^{\mathrm{polyloglog}(d)}\Omega(\eta C)\sum_{\tau=\mathcal{T}_0}^{t-1}(\alpha^{(\tau)})^p. \tag{22}$$

We successively apply Lemma D.9 and $(a-b)^p \geqslant a^p - pb$ for $a \ll b$ to lower bound (22) to obtain:

$$\gamma^{(t)} \geqslant \gamma^{(\mathcal{T}_0)} + e^{\text{polyloglog}(d)}\Omega(\eta C) \sum_{\tau=\mathcal{T}_0}^{t-1} \Big(\frac{\Omega(1)}{C^{2p}\lambda_0^p} - o(\eta)\Big). \tag{23}$$

We apply Induction Hypothesis B.1 in (23) to obtain a bound on $\Gamma^{(t)}$.

$$\begin{aligned} C\Gamma^{(t)} &\geqslant C\Gamma^{(\mathcal{T}_0)} \exp\Big(e^{\text{polyloglog}(d)}\Omega(\eta C) \sum_{\tau=\mathcal{T}_0}^{t-1} \big(\frac{\Omega(1)}{C^{2p}\lambda_0^p} - o(\eta)\big)\Big) \\ &\geqslant \frac{\Omega(C^2)}{D} \exp\Big(e^{\text{polyloglog}(d)}\Omega(\eta C) \sum_{\tau=\mathcal{T}_0}^{t-1} \big(\frac{\Omega(1)}{C^{2p}\lambda_0^p} - o(\eta)\big)\Big). \end{aligned} \tag{24}$$

(24) shows that $\Gamma^{(t)}$ is an non-decreasing sequence. We thus deduce the time $\mathcal{T}_1$ such that $C\Gamma^{(t)} \geqslant \Omega(\lambda_0)/D$.

$$\begin{aligned} &\frac{C^2}{D} \exp\Big(e^{\text{polyloglog}(d)}\Omega(\eta C)\big(\frac{\Omega(1)}{C^{2p}\lambda_0^p} - o(\eta)\big)(\mathcal{T}_1 - \mathcal{T}_0)\Big) = \frac{\lambda_0}{D} \\ &\implies \mathcal{T}_1 = \mathcal{T}_0 + \Theta\Bigg(\frac{C^{2p}\lambda_0^p}{\eta e^{\text{polyloglog}(d)}} \log\Big(\frac{\lambda_0}{C^2}\Big)\Bigg). \end{aligned} \tag{25}$$

We now prove the second part of the lemma. We respectively apply Lemma D.4 and Lemma D.5 to bound $\boldsymbol{\mathcal{S}}^{(t)}$ and $\max_{\tau \leqslant T} |\boldsymbol{\mathcal{N}}^{(\tau)}|$.

$$\boldsymbol{\mathcal{S}}^{(t)} \geqslant \Omega(\lambda_0)^p \quad \text{and} \quad \max_{\tau \leqslant [T]} |\boldsymbol{\mathcal{N}}^{(\tau)}| \leqslant o(\lambda_0)^p. \tag{26}$$

(26) implies for all $t \in [\mathcal{T}_1, T]$, $\boldsymbol{\mathcal{S}}^{(t)} \geqslant \max_{\tau \leqslant [T]} |\boldsymbol{\mathcal{N}}^{(\tau)}|$.

□

D.3.1 Auxiliary lemmas

In this section, we present the auxiliary lemmas needed to prove the main results in subsection D.3.

Lemma D.9. *Let $t \geqslant \mathcal{T}_0$. Then, we always have $\alpha^{(t)} \geqslant \frac{\Omega(1)}{C^2\lambda_0} - o(\eta)$.*

Proof of Lemma D.9. For $t \in [0, \mathcal{T}_0]$, $\alpha^{(t)}$ increases and eventually satisfies $\alpha^{(t)} \geqslant \frac{\Omega(1)}{C^2\lambda_0}$ (Lemma D.2). However, for $t \geqslant \mathcal{T}_0$, $\alpha^{(t)}$ may be non-increasing. Here, we want to quantify the maximum amount of decrease for $t > \mathcal{T}_0$. The worst-case scenario is when $\alpha^{(t)} = \frac{\Omega(1)}{C^2\lambda_0}$. We bound $\alpha^{(t+1)}$ by using Lemma D.4, Lemma D.10 and Lemma D.5.

$$\alpha^{(t+1)} \geqslant \alpha^{(t)} + \Theta(C\eta)(\alpha^{(t)}G^{(t)})^p - \eta(\alpha^{(t)})^p o(\lambda_0)^p. \tag{27}$$

We now apply Induction Hypothesis B.1 in (27) and get:

$$\begin{aligned} \alpha^{(t+1)} &\geqslant \frac{\Omega(1)}{C^2\lambda_0} + \Theta(\eta)\frac{e^{\text{polyloglog}(d)}}{C^{2p-1}\lambda_0^p} - \frac{o(\eta)}{C^{2p}} \\ &\geqslant \frac{\Omega(1)}{C^2\lambda_0} - \frac{o(\eta)}{C^{2p}}. \end{aligned} \tag{28}$$

At time $t+1$, we potentially have $\alpha^{(t+1)} < \frac{\Omega(1)}{C^2\lambda_0}$. In this case, $\alpha^{(t+1)}$ starts to increase again because it is in the range of α's that satisfies Event I (and therefore the update rule in Lemma D.2 holds). Thus, for all $t \geqslant \mathcal{T}_0$, we have $\alpha^{(t)} \geqslant \frac{\Omega(1)}{C^2\lambda_0} - o(\eta)$. □

Lemma D.10. *When the population risk is $\Omega(1)$, we have $\mathbb{E}[\mathfrak{S}(-yF(\boldsymbol{X}))] \geqslant \Omega(1)$.*

Proof of Lemma D.10. Let $(\boldsymbol{X}, y)$ be a data-point. We distinguish two cases:

- $yF(\boldsymbol{X}) > 0$: we apply Lemma K.6 which implies $\mathfrak{S}(-yF(\boldsymbol{X}) \geqslant \log(1 + e^{yF(\boldsymbol{X})})$. Since the population loss is $\Omega(1)$, this implies the aimed result.
- $yF(\boldsymbol{X}) \leqslant 0$: we have necessarily $\mathfrak{S}(-yF(\boldsymbol{X})) \geqslant \Omega(1)$ since the sigmoid function is large for non-positive values.

Therefore, we have $\mathbb{E}[\mathfrak{S}(-yF(\boldsymbol{X}))] \geqslant \Omega(1)$. □

D.4 Event III: $\alpha^{(t)}$ keeps increases again

For $t \in [\mathcal{T}_0, \mathcal{T}_1]$, $\Gamma^{(t)}$ increases until reaching $C\Gamma^{(t)} \geqslant \Omega(\lambda_0)/D$. In this section, we show that this implies that $\alpha^{(t)}$ increases again.

Lemma D.11. *Let $\mathcal{T}_1 = \mathcal{T}_0 + \Theta\left(\frac{\lambda_0^p}{\eta e^{\mathrm{polyloglog}(d)}}\right)$ and $t \in [\mathcal{T}_1, T]$. Since $\boldsymbol{\mathcal{S}}^{(t)} \geqslant \max_{\tau \leqslant T} |\boldsymbol{\mathcal{N}}^{(\tau)}|$, $\alpha^{(t)}$ updates as*

$$\alpha^{(t+1)} = \alpha^{(t)} + \Theta(\eta C)(G^{(t)})^p(\alpha^{(t)})^{p-1}.$$

Consequently, $\alpha^{(t)}$ is non-decreasing until the population risk satisfies $\mathcal{L}(\boldsymbol{A}^{(t)}, \boldsymbol{v}^{(t)}) \leqslant o(1)$. Eventually, $\alpha^{(T)}$ is as large as $\alpha^{(T)} = \frac{\mathrm{polylog}(d)}{C^2\lambda_0}$.

Proof of Lemma D.11. Since $\boldsymbol{\mathcal{S}}^{(t)} \geqslant \max_{\tau \leqslant T} |\boldsymbol{\mathcal{N}}^{(\tau)}|$ (Lemma D.8), the update of $\alpha^{(t)}$ is:

$$\alpha^{(t+1)} - \alpha^{(t)} = \Theta(\eta C)(\alpha^{(t)})^{p-1}(G^{(t)})^p \mathbb{E}[\mathfrak{S}(-yF(\boldsymbol{X}))], \tag{29}$$

Since the population loss is at least $\Omega(1)$ for $t \in [\mathcal{T}_1, T]$, Lemma D.10 implies that $\mathbb{E}[\mathfrak{S}(-yF(\boldsymbol{X}))] \geqslant \Omega(1)$. Besides, we apply Induction Hypothesis B.1 and $\Omega(C) \geqslant 0$ and $e^{\mathrm{polyloglog}(d)} \leqslant \lambda_0$ in (29) and obtain:

$$\begin{cases} \alpha^{(t+1)} \geqslant \alpha^{(t)} + \Theta(\eta C) e^{\mathrm{polyloglog}(d)} (\alpha^{(t)})^{p-1} \\ \alpha^{(t+1)} \leqslant \alpha^{(t)} + \Theta(\eta C) \lambda_0^p (\alpha^{(t)})^{p-1} \end{cases}. \tag{30}$$

(30) and Lemma D.12 show that $\alpha^{(t)}$ increases until reaching $\alpha^{(t)} \leqslant \frac{\mathrm{polylog}(d)}{C^2\lambda_0}$.

□

D.4.1 Auxiliary results

Lemma D.12. *The values of α such that $\mathbb{E}[\mathfrak{S}(-yF(\boldsymbol{X}))] \geqslant \Omega(1)$ is*

$$\alpha \leqslant \frac{\mathrm{polylog}(d)}{C^2\lambda_0}.$$

Proof of Lemma D.12. We say that the sigmoid term is small for a constant κ that satisfies

$$\sum_{\tau=0}^{T} \frac{1}{1 + \exp(\kappa)} \leqslant \mathrm{polylog}(d) \implies \kappa \geqslant \log(T) \iff \kappa \geqslant \mathrm{polylog}(d). \tag{31}$$

Intuitively, (31) means that the sum of the sigmoid terms for all time steps is bounded (up to a logarithmic dependence). In our case, by using (31) and Lemma D.3, the sigmoid $\mathfrak{S}(-yF(\boldsymbol{X}))$ is small when

$$(C^2\alpha\lambda_0)^p \geqslant \mathrm{polylog}(d) \implies \alpha \geqslant \frac{\mathrm{polylog}(d)}{C^2\lambda_0}. \tag{32}$$

□

D.5 Convergence rate of the population loss

Lemma D.13. *Let $t \in [\mathcal{T}_1, T]$. Then, the population loss linearly converges to zero i.e.*

$$\mathcal{L}(\boldsymbol{A}^{(t)}, \boldsymbol{v}^{(t)}) \leqslant \frac{\mathrm{polylog}(d)}{\eta \lambda_0^{2p}(t - \mathcal{T}_1 + 1)} \tag{33}$$

Proof of Lemma D.13. To ease the explanation in this proof, we use the $\tilde{\Omega}$, $\tilde{\Theta}$, $\tilde{O}$ notations to hide the logarithmic dependence. We hide for instance the constant C under this notation. From Lemma D.11, we know that $\alpha^{(t)}$ is lower bounded as:

$$\alpha^{(t+1)} \geqslant \alpha^{(t)} + \Theta(\eta C)(\alpha^{(t)})^{p-1}(G^{(t)})^p \mathbb{E}[\mathfrak{S}(-yF(\boldsymbol{X}))]. \tag{34}$$

Using Lemma D.3, we have $\mathbb{E}[\mathfrak{S}(-yF(\boldsymbol{X}))] \geqslant \mathfrak{S}(-(\alpha^{(t)}G^{(t)})^p)$. Plugging this in (34) yields:

$$\alpha^{(t+1)} \geqslant \alpha^{(t)} + \Theta(\eta) G^{(t)} \frac{(\alpha^{(t)})^{p-1}(G^{(t)})^{p-1}}{1 + \exp((\alpha^{(t)}G^{(t)})^p)}. \tag{35}$$

Since $0 < \alpha^{(t)}G^{(t)} \leqslant \tilde{O}(\lambda_0^{p-1})$,we apply Lemma K.5 and get:

$$\alpha^{(t+1)} \geqslant \alpha^{(t)} + \frac{\tilde{\Omega}(\eta)}{\lambda_0^p} G^{(t)} \log(1 + e^{-(\alpha^{(t)}G^{(t)})^p}). \tag{36}$$

Lemma D.8 implies that $G^{(t)} \geqslant \tilde{\Omega}(\lambda_0)$. Therefore, we have:

$$\alpha^{(t+1)} \geqslant \alpha^{(t)} + \frac{\tilde{\Omega}(\eta)}{\lambda_0^p} \log(1 + e^{-(\alpha^{(t)}G^{(t)})^p}). \tag{37}$$

Let's now assume by contradiction that for $t \in [\mathcal{T}_1, T]$, we have:

$$\log(1 + e^{-(\alpha^{(t)}G^{(t)})^p}) > \frac{\tilde{\Omega}(1)}{\eta \lambda_0^{2p}(t - \mathcal{T}_1 + 1)}. \tag{38}$$

For $t \in [\mathcal{T}_1, T]$, we know that $\alpha^{(t)}G^{(t)}$ is non-decreasing which implies that $(\alpha^{(t)}G^{(t)})^p$ is also non-decreasing. Since $x \mapsto \log(1 + \exp(-x))$ is non-increasing, this implies for $s \leqslant t$ that

$$\frac{\tilde{\Omega}(1)}{\eta \lambda_0^{2p}(t - \mathcal{T}_1)} < \log(1 + e^{-(\alpha^{(t)}G^{(t)})^p}) \leqslant \log(1 + e^{-(\alpha^{(s)}G^{(s)})^p}). \tag{39}$$

Plugging (39) in the update (35) yields for $s \in [\mathcal{T}_1, t]$:

$$\alpha^{(s+1)} > \alpha^{(s)} + \frac{\tilde{\Omega}(1)}{\lambda_0^p(t - \mathcal{T}_1 + 1)}. \tag{40}$$

Let $t \in [\mathcal{T}_1, T]$. We now sum (40) for $s = \mathcal{T}_1, \dots, t$ and obtain:

$$\alpha^{(t+1)} > \alpha^{(\mathcal{T}_1)} + \frac{\tilde{\Omega}(1)(t - \mathcal{T}_1 + 1)}{\lambda_0^p(t - \mathcal{T}_1 + 1)} > \frac{\tilde{\Omega}(1)}{\lambda_0^p}, \tag{41}$$

where we used the fact that $\alpha^{(\mathcal{T}_1)} \geqslant \alpha^{(\mathcal{T}_0)} \geqslant \tilde{\Omega}(1)/(C\lambda_0) > 0$ (Lemma D.2) in the last inequality. Therefore, we have for $t \in [\mathcal{T}_1, T]$, $\alpha^{(t)} \geqslant \tilde{\Omega}(1/\lambda_0^p) > 0$. Let's now show that (41) implies a contradiction. Indeed, we have:

$$\begin{aligned} \eta \lambda_0^{2p}(t - \mathcal{T}_1 + 1) \log(1 + e^{-(\alpha^{(t)}G^{(t)})^p}) &\leqslant \eta \lambda_0^{2p} T \log(1 + e^{-(\alpha^{(t)}G^{(t)})^p}) \\ &\leqslant \eta \lambda_0^{2p} T \log(1 + e^{-\tilde{\Omega}(1)}), \end{aligned} \tag{42}$$

where we used $G^{(t)} \geqslant \Omega(\lambda_0)$ (Lemma D.8) and (41) in the last inequality. We now apply Lemma K.6 and obtain:

$$\eta \lambda_0^{2p}(t - \mathcal{T}_1 + 1) \log(1 + e^{-(\alpha^{(t)}G^{(t)})^p}) \leqslant \frac{\eta \lambda_0^{2p} T}{1 + \exp(\tilde{\Omega}(1))}. \tag{43}$$

Given the values of T, η, λ_0, we finally have:

$$\eta \lambda_0^{2p}(t - \mathcal{T}_1 + 1) \log(1 + e^{-(\alpha^{(t)}G^{(t)})^p}) < \tilde{O}(1), \tag{44}$$

which contradicts (38). Therefore, we obtain the convergence rate:

$$\log(1 + e^{-(\alpha^{(t)}G^{(t)})^p}) \leqslant \frac{\tilde{O}(1)}{\eta \lambda_0^{2p}(t - \mathcal{T}_1 + 1)} \tag{45}$$

We apply Lemma D.14 to bound the left-hand side of (45) and get the aimed result. □

D.5.1 Auxiliary lemmas

Lemma D.14. *Let $t \in [\mathcal{T}_1, T]$. We have:*

$$\Theta(1)\log(1+e^{-(\alpha^{(t)}G^{(t)})^p}) \geqslant \mathcal{L}(\boldsymbol{A}^{(t)}, \boldsymbol{v}^{(t)}).$$

Proof of Lemma D.14. The proof is similar to the one of Lemma D.3. We apply Lemma D.4, Lemma D.8 and Lemma D.5 and get:

$$yF_{\boldsymbol{A}^{(t)},\boldsymbol{v}^{(t)}}(\boldsymbol{X}) \geqslant (\alpha^{(t)})^p\left(C(G^{(t)})^p - o(\lambda_0)^p\right) \geqslant \Theta(1)(\alpha^{(t)}G^{(t)})^p. \tag{46}$$

Using (46), we deduce:

$$\mathbb{E}[\log(1+e^{-yF_{\boldsymbol{A}^{(t)},\boldsymbol{v}^{(t)}}(\boldsymbol{X})})] \leqslant \log(1+e^{-\Theta(1)(\alpha^{(t)}G^{(t)})^p}) \leqslant \Theta(1)\log(1+e^{-(\alpha^{(t)}G^{(t)})^p}), \tag{47}$$

where we applied Lemma K.7 in the last inequality. □

D.6 Fitting the labeling function

We now show that the learner model fits the labeling function.

Lemma D.15. *After T iterations, the population risk converges i.e. $\mathcal{L}(\boldsymbol{A}^{(T)}, \boldsymbol{v}^{(T)}) \leqslant O(1/\mathrm{poly}(d))$. Therefore, $\mathbb{P}_{\mathcal{D}}[f^*(\boldsymbol{X})F_{\boldsymbol{A}^{(t)},\boldsymbol{v}^{(t)}}(\boldsymbol{X}) > 0] \geqslant 1 - o(1)$.*

Proof of Lemma D.15. Since the logistic loss is a surrogate for the 0-1 loss, we have:

$$\mathbb{E}_{\mathcal{D}}[\mathbf{1}_{yF_{\boldsymbol{A}^{(t)},\boldsymbol{v}^{(t)}}(\boldsymbol{X})<0}] \leqslant \mathbb{E}_{\mathcal{D}}[\log(1+e^{-yF_{\boldsymbol{A}^{(t)},\boldsymbol{v}^{(t)}}(\boldsymbol{X})})]. \tag{48}$$

We now apply Lemma D.13 to bound the right-hand side of (48). Given the value of T, we have:

$$\mathbb{P}_{\mathcal{D}}[yF_{\boldsymbol{A}^{(T)},\boldsymbol{v}^{(T)}}(\boldsymbol{X})] = \mathbb{E}_{\mathcal{D}}[\mathbf{1}_{yF_{\boldsymbol{A}^{(T)},\boldsymbol{v}^{(T)}}(\boldsymbol{X})<0}] \leqslant \frac{\mathrm{polylog}(d)}{\eta\lambda_0^{2p}T} \leqslant \frac{1}{\mathrm{poly}(d)}. \tag{49}$$

We now use (49) and Assumption 1 to show that the learner model fits the labeling function. Indeed, we rewrite $\mathbb{P}_{\mathcal{D}}[f^*(\boldsymbol{X})F_{\boldsymbol{A}^{(T)},\boldsymbol{v}^{(T)}}(\boldsymbol{X})]$ as

$$\begin{aligned}\mathbb{P}_{\mathcal{D}}[f^*(\boldsymbol{X})F_{\boldsymbol{A}^{(T)},\boldsymbol{v}^{(T)}}(\boldsymbol{X})] &\geqslant \mathbb{P}_{\mathcal{D}}[yf^*(\boldsymbol{X}) > 0]\mathbb{P}_{\mathcal{D}}[yF_{\boldsymbol{A}^{(T)},\boldsymbol{v}^{(T)}}(\boldsymbol{X}) > 0] \\ &\geqslant (1-d^{-\omega(1)})\left(1-\frac{1}{\mathrm{poly}(d)}\right) \\ &= 1-o(1).\end{aligned} \tag{50}$$

□

D.7 Proof of the induction hypothesis

In this section, we prove Induction Hypothesis B.1.

Proof of Induction Hypothesis B.1. We start by proving that $|\rho^{(t)}| = \Theta(1)$ for all $t \in [T]$. Let $t \in [\mathcal{T}_1, T]$ and $\tau \in [t]$. Using Corollary G.2 and Induction Hypothesis B.1, we upper bound $|\rho^{(\tau)}|$ as:

$$|\rho^{(\tau+1)}| \leqslant |\rho^{(\tau)}| + \frac{\eta\Theta(C)}{D}(\alpha^{(\tau)})^p\left(1+\frac{\lambda_0^{p+1}}{D}\right). \tag{51}$$

Summing (51) for $\tau = 0, \ldots, t-1$ and using $\rho^{(0)} = 0$ lead to

$$|\rho^{(t)}| \leqslant \frac{\eta\Theta(C)}{D}\left(1+\frac{\lambda_0^{p+1}}{D}\right)\sum_{\tau=0}^{T}(\alpha^{(\tau)})^p \tag{52}$$

We now apply Lemma D.16 to bound the sum of $\alpha^{(t)}$'s in (52).

$$|\rho^{(t)}| \leqslant \Theta\left(\frac{1}{D(\alpha^{(0)})^{p-1}(\lambda_0)^p e^{\mathrm{polyloglog}(d)}} + \frac{1}{e^{\mathrm{polyloglog}(d)}}\right). \tag{53}$$

Given the values of the different parameters, (53) implies that $|\rho^{(t)}| \leqslant \Theta(1)$.

We now prove $e^{\gamma^{(t)}} \in [\Omega(1), \lambda_0]$. Since $e^{\gamma^{(t)}}$ is non-decreasing (Corollary G.1), we have $e^{\gamma^{(t)}} \geqslant e^{\gamma^{(0)}} \geqslant \Omega(1)$ for all $t \geqslant 0$. We now prove the upper bound on $e^{\gamma^{(t)}}$. We assume that for all $\tau \leqslant t$, $e^{\gamma^{(\tau)}} \leqslant \lambda_0$. Let's show this inequality for $t+1$. Using Corollary G.1, we have:

$$\begin{aligned} e^{\gamma^{(t+1)}} &\leqslant e^{\gamma^{(t)}} \exp\left(\Theta(C\eta)(\alpha^{(t)})^p \Gamma^{(t)} (G^{(t)})^{p-1}\right) \\ &= \exp\left(\Theta(C\eta) \sum_{\tau=0}^{t} (\alpha^{(\tau)})^p \Gamma^{(\tau)} (G^{(\tau)})^{p-1}\right). \end{aligned} \tag{54}$$

We now apply the induction hypothesis in (54) and get:

$$\begin{aligned} e^{\gamma^{(t+1)}} &\leqslant e^{\gamma^{(t)}} \exp\left(\Theta(C\eta)(\alpha^{(t)})^p \Gamma^{(t)} (G^{(t)})^{p-1}\right) \\ &= \exp\left(\frac{\Theta(\lambda_0^p C\eta)}{D^p} \sum_{\tau=0}^{t} (\alpha^{(\tau)})^p\right) \\ &\leqslant \exp\left(\frac{\Theta(\lambda_0^p C\eta)}{D^p} \sum_{\tau=0}^{T} (\alpha^{(\tau)})^p\right). \end{aligned} \tag{55}$$

We apply Lemma D.16 in (55) and obtain:

$$\begin{aligned} e^{\gamma^{(t+1)}} &\leqslant \exp\left(\Theta\Big(\frac{1}{D^p (\alpha^{(0)})^{p-1} e^{\mathrm{polyloglog}(d)}} + \frac{\lambda_0^p}{D^{p-1} e^{\mathrm{polyloglog}(d)}}\Big)\right) \\ &= \Theta(1) \exp\left(\frac{\lambda_0^p}{D^{p-1} e^{\mathrm{polyloglog}(d)}}\right) \\ &\leqslant \Theta\left(1 + \frac{\lambda_0^p}{D^{p-1} e^{\mathrm{polyloglog}(d)}} + \frac{\lambda_0^{2p}}{D^{2(p-1)} e^{\mathrm{polyloglog}(d)}}\right), \end{aligned} \tag{56}$$

where we used the inequality $e^x \leqslant 1 + x + x^2$ for $x \leqslant 1$ in (56). Given the values of the different parameters, we deduce that $e^{\gamma^{(t+1)}} \leqslant \lambda_0$.

We now prove $e^{\beta^{(t)}} = e^{\mathrm{polyloglog}(d)}$ for $t \in [0, T]$. Since $\beta^{(t)}$ is not updated i.e. $\beta^{(t)} = \beta^{(0)}$ and $\beta^{(0)} = \sigma_{\boldsymbol{M}} = \mathrm{polyloglog}(d)$, we therefore have the aimed result.

Lastly, we prove that $e^{\beta^{(t)}} + (C-1)e^{\gamma^{(t)}} + (D-C)e^{\rho^{(t)}} = \Theta(D)$ for $t \in [T]$. Since $e^{\gamma^{(t)}} \geqslant \Theta(1)$, $e^{\rho^{(t)}} = \Theta(1)$ and $e^{\beta^{(t)}} \geqslant \Theta(1)$, we have:

$$e^{\beta^{(t)}} + (C-1)e^{\gamma^{(t)}} + (D-C)e^{\rho^{(t)}} \geqslant \Theta(D). \tag{57}$$

On the other hand, we have $e^{\gamma^{(t)}} \leqslant \lambda_0$, $e^{\beta^{(t)}} = e^{\mathrm{polyloglog}(d)}$ and $e^{\rho^{(t)}} = \Theta(1)$ which imply:

$$\begin{aligned} e^{\beta^{(t)}} + (C-1)e^{\gamma^{(t)}} + (D-C)e^{\rho^{(t)}} &\leqslant e^{\mathrm{polyloglog}(d)} + (C-1)\lambda_0 + (D-C)\Theta(1) \\ &\leqslant C\lambda_0 + (D-C)\Theta(1) \\ &\leqslant \Theta(D). \end{aligned} \tag{58}$$

□

D.7.1 Auxiliary lemmas

Lemma D.16. *The sum of the $\alpha^{(t)}$'s is bounded as:*

$$\sum_{\tau=0}^{T} (\alpha^{(\tau)})^p = \Theta\left(\frac{1}{\eta C (\alpha^{(0)})^{p-1} (\lambda_0)^p e^{\mathrm{polyloglog}(d)}} + \frac{D}{C\eta e^{\mathrm{polyloglog}(d)}}\right).$$

Proof of Lemma D.16. We first decompose the sum of $\alpha^{(t)}$'s.

$$\sum_{\tau=0}^{T}(\alpha^{(\tau)})^p = \sum_{\tau=0}^{\mathcal{T}_0-1}(\alpha^{(\tau)})^p + \sum_{\tau=\mathcal{T}_0}^{\mathcal{T}_1-1}(\alpha^{(\tau)})^p + \sum_{\tau=\mathcal{T}_1}^{T}(\alpha^{(\tau)})^p. \tag{59}$$

We apply Lemma D.2 and Lemma D.8 to rewrite (59).

$$\sum_{\tau=0}^{T}(\alpha^{(\tau)})^p = \frac{\Theta(1)}{\eta C(\alpha^{(0)})^{p-1}(\lambda_0)^p e^{\mathrm{polyloglog}(d)}} + \frac{\log(\lambda_0)}{\eta e^{\mathrm{polyloglog}(d)}} + \sum_{\tau=\mathcal{T}_1}^{T}(\alpha^{(\tau)})^p. \tag{60}$$

Now, we aim to obtain the value of the last summand in (60). Using Corollary G.1, we have

$$C\Gamma^{(T)} \geqslant C\Gamma^{(\mathcal{T}_1)}\exp\left(\frac{C\eta}{D}e^{\mathrm{polyloglog}(d)}\sum_{\tau=\mathcal{T}_1}^{T}(\alpha^{(\tau)})^p\right). \tag{61}$$

We finally apply Lemma D.8 and Induction Hypothesis B.1 in (61) to get:

$$\sum_{\tau=\mathcal{T}_1}^{T}(\alpha^{(\tau)})^p \leqslant \frac{\Theta(D)}{C\eta e^{\mathrm{polyloglog}(d)}}. \tag{62}$$

To obtain the aimed result, we plug (62) in (60) and use $\frac{\log(\lambda_0)}{\eta e^{\mathrm{polyloglog}(d)}} \leqslant \frac{\Theta(D)}{C\eta e^{\mathrm{polyloglog}(d)}}$. □

E From idealized to real learning process

In Appendix D, we analyzed the ideal learning process. We now aim to bridge the gap between the idealized and realistic cases. Given our initialization, $\widehat{\boldsymbol{v}}^{(t)}$ has a component in $\mathrm{span}(\boldsymbol{w}^*)^\perp$ i.e.

$$\widehat{\boldsymbol{v}}^{(t)} = \widehat{\alpha}^{(t)}\boldsymbol{w}^* + \varepsilon_{\boldsymbol{v}}^{(t)}\boldsymbol{u}^{(t)}$$

where $\boldsymbol{u}^{(t)} \in \mathbb{R}^d$ such that $\boldsymbol{u}^{(t)} \perp \boldsymbol{w}^*$ and $\|\boldsymbol{u}^{(t)}\|_2 = 1$. Thus, one main difference between the two cases is that we initialize $\alpha^{(0)} \geqslant \nu^{1/(p-1)}$ in the idealized case while $\widehat{\alpha}^{(0)} \leqslant \omega$. Thus, the proof strategy consists in i) $t \in [0, \mathscr{T}]$, $\widehat{\alpha}^{(t)}$ increases until having $\widehat{\alpha}^{(t)} \geqslant \nu^{1/(p-1)}$ (subsubsection E.1.1) while $\varepsilon_{\boldsymbol{v}}^{(t)}$ (subsubsection E.1.2) and $\widehat{A}_{i,j}^{(t)}$ (subsubsection E.1.3) stay tiny. ii) $t \in [\mathscr{T}, T]$, compare the realistic and idealized iterates. We remind the GD update of $\widehat{\alpha}^{(t)}$.

$$\begin{aligned}
&\widehat{\alpha}^{(t+1)} - \widehat{\alpha}^{(t)} \\
&= \eta\underbrace{\frac{D}{N}\sum_{i=1}^{N} y[i]\mathfrak{S}\big(-y[i]F(\boldsymbol{X}[i])\big)\sum_{j\in\mathcal{S}_{\ell(\boldsymbol{X}[i])}}\sigma'\Big(\sum_{k=1}^{D}\widehat{S}_{j,k}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)},\boldsymbol{X}_k[i]\rangle\Big)\sum_{r=1}^{D}\widehat{S}_{j,r}^{(t)}\langle\boldsymbol{w}^*,\boldsymbol{X}_r[i]\rangle}_{\widehat{\boldsymbol{\mathcal{S}}}^{(t)}} \\
&+ \eta\underbrace{\frac{D}{N}\sum_{i=1}^{N} y[i]\mathfrak{S}\big(-y[i]F(\boldsymbol{X}[i])\big)\sigma'\Big(\sum_{k=1}^{D}\widehat{S}_{j,k}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)},\boldsymbol{X}_k[i]\rangle\Big)\sum_{r=1}^{D}\widehat{S}_{j,r}^{(t)}\langle\boldsymbol{w}^*,\boldsymbol{X}_r[i]\rangle}_{\widehat{\boldsymbol{\mathcal{N}}}^{(t)}}.
\end{aligned} \tag{GD-$\widehat{\alpha}$}$$

E.1 Bound on the iterates during the initial steps ($t \in [0, \mathscr{T}]$)

Since we randomly initialize $\widehat{\boldsymbol{v}}^{(0)}$ with tiny variance, we need to take into account the linear part of the activation function. Lemma E.8 shows that we can overlook the power part of the activation and consider $\sigma(x) = \nu x$ as long as $\widehat{\alpha}^{(t)} \geqslant \nu^{1/(p-1)}$.

E.1.1 $\widehat{\alpha}^{(t)}$ initially increases

Lemma E.1. *Let $\mathscr{T} = \Theta\Big(\frac{1}{\eta\nu^{(p-2)/(p-1)}e^{\mathrm{polyloglog}(d)}}\Big)$. For all $t \in [0, \mathscr{T}]$, $\widehat{\alpha}^{(t)}$ is updated as*

$$\widehat{\alpha}^{(t+1)} = \widehat{\alpha}^{(t)} + \Theta(\eta\nu)e^{\beta}.$$

Consequently, $\widehat{\alpha}^{(t)}$ is non-decreasing and after $\mathscr{T}$ iterations, we have $\widehat{\alpha}^{(t)} \geqslant \nu^{1/(p-1)}$ for $t \geqslant \mathscr{T}$.

Proof of Lemma E.1. Let $t \geqslant 0$. We apply Lemma E.5 and Lemma E.6 to respectively bound $\widehat{\boldsymbol{\mathcal{S}}}^{(t)}$ and $\widehat{\boldsymbol{\mathcal{N}}}^{(t)}$ in the update of $\widehat{\alpha}^{(t)}$.

$$\widehat{\alpha}^{(t+1)} = \widehat{\alpha}^{(t)} + \Theta(C\eta\nu)e^{\beta}. \tag{63}$$

(63) indicates that $\widehat{\alpha}^{(t)}$ is a non-decreasing sequence. Therefore, there exists a time $\mathscr{T}$ such that $\widehat{\alpha}^{(\mathscr{T})} = \nu^{1/(p-1)}$. Summing (63) for $t = 0, \ldots, \mathscr{T} - 1$ yields $\mathscr{T} = \Theta\Big(\frac{1}{\eta\nu^{(p-2)/(p-1)}e^{\beta}}\Big)$. □

E.1.2 Bound on $\varepsilon_{\boldsymbol{v}}$

We now show that for $t \in [0, \mathscr{T}]$, the orthogonal component $\varepsilon_{\boldsymbol{v}}^{(t)}$ stays small.

Lemma E.2. *Assume that we run GD on the empirical risk (E) for T iterations with parameters set as in Parametrization 3.1. For $t \in [0, \mathscr{T}]$, the orthogonal component $\varepsilon_{\boldsymbol{v}}$ satisfies*

$$\varepsilon_{\boldsymbol{v}}^{(t+1)} \leqslant \Big(1 + \eta\nu\frac{\mathrm{poly}(D)}{\sqrt{d}}\Big)\varepsilon_{\boldsymbol{v}}^{(t)} + \eta\zeta,$$

where $\zeta = \frac{\mathrm{poly}(D)}{\sqrt{N}}$.

Proof of Lemma E.2. Let $\mathbf{P} = (\mathbf{I} - \boldsymbol{w}^*\boldsymbol{w}^{*\top})$ and $\mathring{\boldsymbol{v}}^{(t)} = \widehat{\alpha}^{(t)}\boldsymbol{w}^*$. The projected update of $\widehat{\boldsymbol{v}}$ satisfies:

$$\begin{aligned}
&\|\mathbf{P}\widehat{\boldsymbol{v}}^{(t+1)} - \mathbf{P}\widehat{\boldsymbol{v}}^{(t)}\|_2 \\
\leqslant& D\eta\nu\Bigg\|\frac{1}{N}\sum_{i=1}^{N} y[i]\mathfrak{S}(-y[i]F_{\widehat{\boldsymbol{v}}}(\boldsymbol{X}[i]))\sum_{m=1}^{D}\sum_{b=1}^{D}\widehat{\boldsymbol{S}}_{m,b}^{(t)}\mathbf{P}\boldsymbol{X}_b[i] && (64)\\
&- \mathbb{E}\Big[y\mathfrak{S}(-yF_{\widehat{\boldsymbol{v}}}(\boldsymbol{X}))\sum_{m=1}^{D}\sum_{b=1}^{D}\widehat{\boldsymbol{S}}_{m,b}^{(t)}\mathbf{P}\boldsymbol{X}_b\Big]\Bigg\|_2 && (65)\\
&+ D\eta\nu\Bigg\|\mathbb{E}\Big[y\mathfrak{S}(-yF_{\widehat{\boldsymbol{v}}}(\boldsymbol{X}))\sum_{m=1}^{D}\sum_{b=1}^{D}\widehat{\boldsymbol{S}}_{m,b}^{(t)}\mathbf{P}\boldsymbol{X}_b\Big] && (66)\\
&- \mathbb{E}\Big[y\mathfrak{S}(-yF_{\mathring{\boldsymbol{v}}}(\boldsymbol{X}))\sum_{m=1}^{D}\sum_{b=1}^{D}\widehat{\boldsymbol{S}}_{m,b}^{(t)}\mathbf{P}\boldsymbol{X}_b\Big]\Bigg\|_2. && (67)
\end{aligned}$$

Summand 1: $\|(64) - (65)\|_2$. Using the matrix Hoeffding inequality, we have with high probability, $\|(80) - (81)\|_2 \leqslant 16\sqrt{\log(d)\sum_{i=1}^{N} M_i^2}$, where $\Big\|\frac{\eta\nu D y[i]}{N}\mathfrak{S}(-y[i]F_{\widehat{\boldsymbol{v}}}(\boldsymbol{X}^{(i)}))\sum_{m=1}^{D}\sum_{b=1}^{D}\widehat{\boldsymbol{S}}_{m,b}^{(t)}\mathbf{P}\boldsymbol{X}_b^{(i)}\Big\|_2^2 \leqslant M_i^2$. Induction Hypothesis B.1 and $\|\mathbf{P}\boldsymbol{X}_u^{(i)}\|_2 \leqslant \sigma^2 d\log(d)$ imply $M_i^2 \leqslant \frac{\eta^2\nu^2\mathrm{poly}(D)}{N^2}$. We deduce that $\|(64) - (65)\|_2 \leqslant \eta\nu\frac{\mathrm{poly}(D)}{\sqrt{N}}$.

Summand 2: $\|(66) - (67)\|_2$. We use the 1-Lipschitzness of the sigmoid function and get:

$$\begin{aligned}
&\|(66) - (67)\|_2 \\
\leqslant& \eta\nu\Bigg\|\mathbb{E}\Bigg[Dy\sum_{m=1}^{D}\sum_{b=1}^{D}\widehat{\boldsymbol{S}}_{m,b}^{(t)}\mathbf{P}\boldsymbol{X}_u\cdot \\
&\Big[\mathfrak{S}\Big(-y\nu D\sum_{a=1}^{D}\sum_{a'=1}^{D}\widehat{\boldsymbol{S}}_{a,a'}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_{a'}\rangle\Big) - \mathfrak{S}\Big(-y\nu D\sum_{a=1}^{D}\sum_{a'=1}^{D}\widehat{\boldsymbol{S}}_{a,a'}^{(t)}\langle\mathring{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_{a'}\rangle\Big)\Big]\Bigg]\Bigg\|_2 \\
\leqslant& \Theta(\eta\nu D)\mathbb{E}\Big[\sum_{m=1}^{D}\sum_{r=1}^{D}\widehat{\boldsymbol{S}}_{m,r}^{(t)}\sum_{a=1}^{D}\Big|D\sum_{a'=1}^{D}\widehat{\boldsymbol{S}}_{a,a'}^{(t)}\langle\mathring{\boldsymbol{v}}^{(t)} - \widehat{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_{a'}\rangle\Big|\sum_{b=1}^{D}\widehat{\boldsymbol{S}}_{m,u}^{(t)}\|\mathbf{P}\boldsymbol{X}_b\|_2\Big] \\
\leqslant& \Theta(\eta\nu D)\mathbb{E}\Big[\sum_{m=1}^{D}\lambda_0\varepsilon_{\boldsymbol{v}}^{(t)}\cdot\Big|\langle\boldsymbol{u}^{(t)}, \sum_{a'=1}^{D}\boldsymbol{\xi}_{a'}\rangle\Big|\cdot\sum_{b=1}^{D}\lambda_0\|\boldsymbol{\xi}_b\|_2\Big]. && (68)
\end{aligned}$$

where we applied Induction Hypothesis B.1 in (68). Since with high probability, $\|\boldsymbol{\xi}_b\|_2 \leqslant \sigma\sqrt{d\log(d)}$, $\left|\langle \boldsymbol{u}^{(t)}, \sum_{r=1}^{D} \boldsymbol{\xi}_r\rangle\right| \leqslant \sqrt{D\log(d)}\sigma$, we finally have:

$$\|(66)-(67)\|_2 \leqslant \eta\nu \cdot \mathrm{poly}(D)\varepsilon_{\boldsymbol{v}}^{(t)}\sigma = \eta\nu\frac{\mathrm{poly}(D)}{\sqrt{d}}\varepsilon_{\boldsymbol{v}}^{(t)}.$$

Combining the bounds on Summands 1 and 2 yields the aimed result. □

We now use Lemma E.2 to show that $\varepsilon_{\boldsymbol{v}}$ stays small.

Lemma E.3. *For all $t \leqslant \mathscr{T}$, $\varepsilon_{\boldsymbol{v}}^{(t)} \leqslant \omega\sqrt{d\log(d)} + \nu^{1/(p-1)}\frac{\mathrm{poly}(D)}{\sqrt{N}}$. By setting $N = \mathrm{poly}(d)$, we have: $\varepsilon_{\boldsymbol{v}}^{(t)} \leqslant 1/\mathrm{poly}(d)$.*

Proof of Lemma E.3. Unraveling Lemma E.2 for $t = 0, \dots, \mathscr{T}$ and using $\varepsilon_{\boldsymbol{v}}^{(0)} \leqslant \omega\sqrt{d\log(d)}$ (with high probability) leads to:

$$\varepsilon_{\boldsymbol{v}}^{(\mathscr{T})} \leqslant \omega\sqrt{d\log(d)} + \eta\nu\zeta\frac{\left(1+\eta\nu\frac{\mathrm{poly}(D)}{\sqrt{d}}\right)^{\mathscr{T}} - 1}{\eta\nu\frac{\mathrm{poly}(D)}{\sqrt{d}}} \leqslant \omega\sqrt{d\log(d)} + 2\mathscr{T}\eta\nu\zeta, \tag{69}$$

where we used $(1+x)^y \leqslant 1 + 2yx$ for $x \ll 1$ and $y \geqslant 0$. Plugging the value of $\mathscr{T}$ in (69) yields the aimed result. □

E.1.3 $\widehat{A}_{a,b}^{(t)}$ stays small

We finally show that $\widehat{A}_{a,b}^{(t)}$ remains tiny for $t \in [0, \mathscr{T}]$.

Lemma E.4. *Let $a, b \in [D]$. We have $|\widehat{A}_{a,b}^{(t)}| \leqslant \omega\sqrt{d\log(d)} + \frac{\Theta(\nu^{2/(p-1)})}{De^{\beta}}$.*

Proof of Lemma E.4. We remind that the GD update of $\widehat{A}_{a,b}^{(t)}$ is

$$\widehat{A}_{a,b}^{(t+1)} = \widehat{A}_{a,b}^{(t)} + \frac{\nu\eta}{N}\sum_{i=1}^{N}\mathfrak{S}\big(-y[i]F(\boldsymbol{X}[i])\big)\widehat{S}_{a,b}^{(t)}\sum_{m\neq b}\widehat{S}_{a,m}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_m[i] - \boldsymbol{X}_b[i]\rangle. \tag{70}$$

The proof is by induction. We assume that $\widehat{A}_{a,b}^{(t)} \leqslant \omega\sqrt{d\log(d)} + \frac{\Theta(\nu^{2/(p-1)})}{De^{\beta}}$. We first apply Cauchy-Schwarz on (70) and get:

$$|\widehat{A}_{a,b}^{(t+1)}| \leqslant |\widehat{A}_{a,b}^{(t)}| + \nu\eta\widehat{S}_{a,b}^{(t)}\sum_{m\neq b}\widehat{S}_{a,m}^{(t)}\|\widehat{\boldsymbol{v}}^{(t)}\|_2. \tag{71}$$

Using the induction hypothesis, we have $\widehat{S}_{a,b}^{(t)} \leqslant \Theta(1)/D$. Thus, we have

$$|\widehat{A}_{a,b}^{(t+1)}| \leqslant |\widehat{A}_{a,b}^{(t)}| + \frac{\Theta(\nu\eta)}{D}\|\widehat{\boldsymbol{v}}^{(t)}\|_2 \leqslant |\widehat{A}_{a,b}^{(t)}| + \frac{\Theta(\nu\eta)}{D}\sqrt{(\widehat{\alpha}^{(t)})^2 + (\varepsilon_{\boldsymbol{v}}^{(t)})^2}. \tag{72}$$

We sum (72) and get:

$$|\widehat{A}_{a,b}^{(\mathscr{T})}| \leqslant |\widehat{A}_{a,b}^{(0)}| + \frac{\Theta(\nu\eta)}{D}\sum_{t=0}^{\mathscr{T}-1}\sqrt{(\widehat{\alpha}^{(t)})^2 + (\varepsilon_{\boldsymbol{v}}^{(t)})^2}. \tag{73}$$

We now use Lemma E.1 and Lemma E.3 in (73) and get:

$$\begin{aligned}|\widehat{A}_{a,b}^{(\mathscr{T})}| &\leqslant \omega\sqrt{d\log(d)} + \frac{\Theta(\nu^{1/(p-1)})}{De^{\beta}}\sqrt{\nu^{2/(p-1)} + \left(\omega\sqrt{d\log(d)} + \nu^{1/(p-1)}\frac{\mathrm{poly}(D)}{\sqrt{N}}\right)^2} \\ &\leqslant \omega\sqrt{d\log(d)} + \frac{\Theta(\nu^{2/(p-1)})}{De^{\beta}}.\end{aligned} \tag{74}$$

□

E.1.4 Auxiliary lemmas

Lemma E.5. *Let $\mathscr{T}$ be the time where $\widehat{\alpha}^{(t)} \geqslant \nu^{1/(p-1)}$. Let $t \in [0, \mathscr{T}]$, $i \in [N]$ and $j \in \mathcal{S}_{\ell(\boldsymbol{X}[i])}$. We have:*

$$y[i]\langle \boldsymbol{w}^*, \boldsymbol{O}_j^{(t)}[i]\rangle = e^{\beta}.$$

This implies $\widehat{\boldsymbol{\mathcal{S}}}^{(t)} = Ce^{\beta}$ for all $t \in [0, \mathscr{T}]$.

Proof of Lemma E.5. We successively apply Lemma E.4 and Lemma K.3 to get:

$$\begin{aligned} y[i]\langle \boldsymbol{w}^*, \boldsymbol{O}_j^{(t)}[i]\rangle &= D\Bigg(\widehat{S}_{j,j}^{(t)} + \sum_{k\in\mathcal{S}_{\ell(\boldsymbol{X}[i])}} \widehat{S}_{j,k}^{(t)} + y[i] \sum_{h\neq \ell(\boldsymbol{X}[i])} \sum_{r\in\mathcal{S}_h} \widehat{S}_{j,r}^{(t)}\delta_r\Bigg) \\ &\geqslant \Theta\Bigg(e^{\mathrm{polyloglog}(d)} + \Theta(C-1) - \Theta(1) \sum_{h\neq \ell(\boldsymbol{X}[i])} \sum_{r\in\mathcal{S}_h} \delta_{h,r}\Bigg) \qquad (75) \\ &\geqslant \Theta\Big(e^{\mathrm{polyloglog}(d)} + \Theta(C-1) - \Theta(qD\log(d))\Big) \qquad (76) \\ &\geqslant e^{\mathrm{polyloglog}(d)}. \end{aligned}$$

Similarly, we also have $y[i]\langle \boldsymbol{w}^*, \boldsymbol{O}_j^{(t)}[i]\rangle \leqslant e^{\mathrm{polyloglog}(d)}$. □

Lemma E.6. *Let $t \in [0, \mathscr{T}]$. Assume that $N = \mathrm{poly}(d)$. With high probability, $\widehat{\boldsymbol{\mathcal{N}}}^{(t)} \leqslant 1/\mathrm{poly}(d)$.*

Proof of Lemma E.6. During this time phase, the sigmoid stays constant. Therefore, we have $\mathbb{E}[\widehat{\boldsymbol{\mathcal{N}}}^{(t)}] = 0$. Therefore, we apply Hoeffding inequality and Lemma D.6 to get:

$$\widehat{\boldsymbol{\mathcal{N}}}^{(t)} \leqslant \sqrt{8N\log(d)\max_{i\in[N]} M_i^2}, \qquad (77)$$

where A_i is a constant such that $\left|\frac{1}{N}\sum_{j\notin\mathcal{S}_{\ell(\boldsymbol{X}[i])}} y[i]\langle \boldsymbol{w}^*, \boldsymbol{O}_j^{(t)}[i]\rangle\right| \leqslant M_i \leqslant \frac{\mathrm{poly}(D)}{N^2}$. Since $N = \mathrm{poly}(d)$, we finally proved $\widehat{\boldsymbol{\mathcal{N}}}^{(t)} \leqslant 1/\mathrm{poly}(d)$. □

Lemma E.7. *Let $t \leqslant \mathscr{T}$. The sum of $\widehat{\alpha}^{(t)}$'s is bounded as:*

$$\sum_{\tau=0}^{t} \widehat{\alpha}^{(\tau)} = t\widehat{\alpha}^{(0)} + \Theta(\eta\nu)e^{\beta}t^2. \qquad (78)$$

Proof of Lemma E.7. Let $\tau \in [0, \mathscr{T}]$. We sum the update rule of $\widehat{\alpha}^{(t)}$ (Lemma E.1) and obtain: $\widehat{\alpha}^{(\tau)} = \widehat{\alpha}^{(0)} + \Theta(C\eta\nu)e^{\beta}\tau$. Summing again this update yields the aimed result.

□

Lemma E.8. *Let $(\boldsymbol{X}, \cdot) \sim \mathcal{D}$ and $j \in [D]$. Assume that $\widehat{\alpha}^{(t)} \leqslant \nu^{1/(p-1)}$. Then, we have:*

$$\sigma'\Big(D\sum_{k=1}^{D} \widehat{S}_{j,k}^{(t)}\langle \widehat{\boldsymbol{v}}, \boldsymbol{X}_k\rangle\Big) = \begin{cases} \Theta(\nu) & \text{if } \widehat{\alpha}^{(t)} \leqslant \nu^{1/(p-1)} \\ \Theta(p)\Big(D\sum_{k=1}^{D} \widehat{S}_{j,k}^{(t)}\langle \widehat{\boldsymbol{v}}, \boldsymbol{X}_k\rangle\Big)^{p-1} & \text{otherwise} \end{cases}.$$

Proof of Lemma E.8. We remind that the derivative of the activation function $\sigma'(x) = px^{p-1} + \nu$. We first remark that for all x, $\sigma'(x) \geqslant \nu$. Besides, we have since $p-1$ is even,

$$px^{p-1} + \nu \leqslant 3\nu \iff |x| \leqslant \left(\frac{2\nu}{p}\right)^{1/(p-1)}. \qquad (79)$$

In our case, we have $x = D\widehat{\alpha}^{(t)}\sum_{k=1}^{D} \widehat{S}_{j,k}^{(t)}\langle \boldsymbol{w}^*, \boldsymbol{X}_k\rangle$. Using Induction Hypothesis B.1, we have $|x| \geqslant e^{\mathrm{polyloglog}(d)}\alpha^{(t)}$. Therefore, a sufficient condition for (79) to hold is $\widehat{\alpha}^{(t)} \leqslant \nu^{1/(p-1)}e^{-\mathrm{polyloglog}(d)}$. Since $e^{\mathrm{polyloglog}(d)} \ll \mathrm{poly}(d)$, we can simplify this condition as $\widehat{\alpha}^{(t)} \leqslant \nu^{1/(p-1)}$. Proving the second part of the lemma can be done as in the proof of Lemma D.7. □

E.2 Coupling between the semi-idealized and realistic processes ($t \in [\mathscr{T}, T]$)

In this section, we aim to bound the realistic iterates $\widehat{A}_{i,j}^{(t)}$ and $\widehat{\boldsymbol{v}}^{(t)}$ for $t \in [\mathscr{T}, T]$. For this reason, we introduce a "semi-idealized" learning process (subsubsection E.2.1) which may be viewed as a mid-point between the idealized and realistic process. We first bound the iterates in this process. Then, using this process, we show that $\varepsilon_{\boldsymbol{v}}^{(t)}$ (subsubsection E.2.2) and $\Delta_{\boldsymbol{A}}^{(t)} := \max_{i \neq j} |\widehat{A}_{i,j}^{(t)} - \breve{A}_{i,j}^{(t)}|$ (subsubsection E.2.4) stay small. Here, $\breve{A}_{i,j}^{(t)}$ is the semi-idealized attention matrix coefficient. Finally, since $\varepsilon_{\boldsymbol{v}}^{(T)}$ and $\Delta_{\boldsymbol{A}}^{(T)}$ are small, the final iterates $\widehat{\alpha}^{(T)}$ and $\alpha^{(T)}$ are equal (subsubsection E.2.6) and thus, the model fits the labeling function (subsubsection E.2.7).

E.2.1 Defining the semi-idealized process

We define an intermediate learning process that we refer to as the "semi-idealized" process. This process starts at time $t = \mathscr{T}$ involves two parameters: the semi-idealized value vector $\breve{\boldsymbol{v}}$ and semi-idealized attention matrix $\breve{\boldsymbol{A}}$ defined as

- the value vector $\breve{\boldsymbol{v}}$ is fixed and satisfies $\breve{\boldsymbol{v}}^{(t-\mathscr{T})} = \widehat{\alpha}^{(t)} \boldsymbol{w}^*$ for $t \in [\mathscr{T}, T]$.
- $\breve{A}_{i,j}^{(t-\mathscr{T})}$ is a trainable parameter and is initialized as $\breve{A}_{i,j}^{(0)} = 0$ for $i \neq j$.

Therefore, the only trainable parameter in this process is $\breve{\boldsymbol{A}}$. In the semi-idealized process, we minimize the population risk

$$\min_{\breve{\boldsymbol{A}}} \ \mathbb{E}_{\mathcal{D}}\big[\log\big(1 + e^{-yF(\boldsymbol{X})}\big)\big] := \breve{\mathcal{L}}(\breve{\boldsymbol{v}}, \breve{\boldsymbol{A}}). \tag{$\tilde{\text{P}}$}$$

We remark that such process present similarities to the idealized case. In particular, it satisfies all the invariance and symmetry properties from Lemma 4.1. We thus define

- $\breve{A}_{i,j}^{(t)} = \breve{\gamma}^{(t)}$ for all $\ell \in [L]$ and $i, j \in \mathcal{S}_\ell$.
- $\breve{A}_{i,j}^{(t)} = \breve{\rho}^{(t)}$ for all $\ell, m \in [L]$ such that $m \neq \ell$ and $i \in \mathcal{S}_\ell$ and $j \in \mathcal{S}_m$.

Therefore, $\breve{\gamma}^{(t)}$ and $\breve{\rho}^{(t)}$ are respectively updated as in Lemma G.1 and Lemma G.2. We define also the softmax terms

$$\breve{\Lambda}^{(t)} = \frac{e^{\beta}}{e^{\beta} + (C-1)e^{\breve{\gamma}^{(t)}} + (D-C)e^{\breve{\rho}^{(t)}}}, \qquad \breve{\Gamma}^{(t)} = \frac{e^{\breve{\gamma}^{(t)}}}{e^{\beta} + (C-1)e^{\breve{\gamma}^{(t)}} + (D-C)e^{\breve{\rho}^{(t)}}},$$
$$\breve{\Xi}^{(t)} = \frac{e^{\breve{\rho}^{(t)}}}{e^{\beta} + (C-1)e^{\breve{\gamma}^{(t)}} + (D-C)e^{\breve{\rho}^{(t)}}}, \qquad \breve{G}^{(t)} = D(\breve{\Lambda}^{(t)} + (C-1)\breve{\Gamma}^{(t)}).$$

We finally assume Induction Hypothesis B.1 for this process. This latter can be proved using the same arguments as in subsection D.7.

E.2.2 Realistic dynamics are mainly on $\operatorname{span}(\boldsymbol{w}^*)$

We previously showed in Lemma E.3 that $\varepsilon_{\boldsymbol{v}}^{(t)}$ is small in the initial steps. We now show that it stays small during the whole process.

Lemma E.9. *Assume that we run GD on the empirical risk (E) for T iterations with parameters set as in Parametrization 3.1 and the number of samples is $N = \operatorname{poly}(d)$. Then, $\widehat{\boldsymbol{v}}^{(t)}$ mainly lies in $\operatorname{span}(\boldsymbol{w}^*)$ i.e. for $t \leqslant T$, $\varepsilon_{\boldsymbol{v}}^{(t)} \leqslant 1/\operatorname{poly}(d)$.*

We now proceed to the proof of Lemma E.9. We first characterize the recursion satisfied by $\varepsilon_{\boldsymbol{v}}^{(t)}$.

Lemma E.10. *Assume that we run GD on the empirical risk (E) for T iterations with parameters set as in Parametrization 3.1. Then, $\varepsilon_{\boldsymbol{v}}^{(t)}$ satisfies for $t \in [\mathscr{T}, T]$*

$$\varepsilon_{\boldsymbol{v}}^{(t+1)} \leqslant \big(1 + \eta(\widehat{\alpha}^{(t)})^{p-1}\frac{\operatorname{poly}(D)}{\sqrt{d}}\big)\varepsilon_{\boldsymbol{v}}^{(t)} + \eta\zeta, \qquad \textit{where } \zeta = \frac{\operatorname{poly}(D)}{\sqrt{N}}.$$

Proof of Lemma E.10. Let $\mathbf{P} := (\mathbf{I} - \boldsymbol{w}^*\boldsymbol{w}^{*\top})$, $\mathring{\boldsymbol{v}}^{(t)} := \widehat{\alpha}^{(t)}\boldsymbol{w}^*$ and $t \in [\mathscr{T}, T]$. The projected update of $\widehat{\boldsymbol{v}}$ satisfies:

$$
\begin{aligned}
&\|\mathbf{P}\widehat{\boldsymbol{v}}^{(t+1)} - \mathbf{P}\widehat{\boldsymbol{v}}^{(t)}\|_2 \\
\leqslant &\eta\Bigg\|\frac{1}{N}\sum_{i=1}^{N} Dy[i]\mathfrak{S}(-y[i]F_{\widehat{\boldsymbol{v}}}(\boldsymbol{X}[i]))\sum_{m=1}^{D}\left(D\sum_{r=1}^{D}\widehat{S}_{m,r}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_r^{(i)}\rangle\right)^{p-1}\sum_{b=1}^{D}\widehat{S}_{m,b}^{(t)}\mathbf{P}\boldsymbol{X}_b[i] && (80)\\
&-\mathbb{E}\Bigg[Dy\mathfrak{S}(-yF_{\widehat{\boldsymbol{v}}}(\boldsymbol{X}))\sum_{m=1}^{D}\left(D\sum_{r=1}^{D}\widehat{S}_{m,r}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_r\rangle\right)^{p-1}\sum_{b=1}^{D}\widehat{S}_{m,b}^{(t)}\mathbf{P}\boldsymbol{X}_b\Bigg]\Bigg\|_2 && (81)\\
&+\eta\Bigg\|\mathbb{E}\Bigg[Dy\mathfrak{S}(-yF_{\widehat{\boldsymbol{v}}}(\boldsymbol{X}))\sum_{m=1}^{D}\left(D\sum_{r=1}^{D}\widehat{S}_{m,r}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_r\rangle\right)^{p-1}\sum_{b=1}^{D}\widehat{S}_{m,b}^{(t)}\mathbf{P}\boldsymbol{X}_b\Bigg] && (82)\\
&-\mathbb{E}\Bigg[Dy\mathfrak{S}(-yF_{\widehat{\boldsymbol{v}}}(\boldsymbol{X}))\sum_{m=1}^{D}\left(D\sum_{r=1}^{D}\widehat{S}_{m,r}^{(t)}\langle\mathring{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_r\rangle\right)^{p-1}\sum_{b=1}^{D}\widehat{S}_{m,b}^{(t)}\mathbf{P}\boldsymbol{X}_b\Bigg]\Bigg\|_2 && (83)\\
&+\eta\Bigg\|\mathbb{E}\Bigg[Dy\mathfrak{S}(-yF_{\widehat{\boldsymbol{v}}}(\boldsymbol{X}))\sum_{m=1}^{D}\left(D\sum_{r=1}^{D}\widehat{S}_{m,r}^{(t)}\langle\mathring{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_r\rangle\right)^{p-1}\sum_{b=1}^{D}\widehat{S}_{m,b}^{(t)}\mathbf{P}\boldsymbol{X}_b\Bigg] && (84)\\
&-\mathbb{E}\Bigg[Dy\mathfrak{S}(-yF_{\mathring{\boldsymbol{v}}}(\boldsymbol{X}))\sum_{m=1}^{D}\left(D\sum_{r=1}^{D}\widehat{S}_{m,r}^{(t)}\langle\mathring{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_r\rangle\right)^{p-1}\sum_{b=1}^{D}\widehat{S}_{m,b}^{(t)}\mathbf{P}\boldsymbol{X}_b\Bigg]\Bigg\|_2. && (85)
\end{aligned}
$$

Remark that (85) is equal to zero because $\mathbb{E}[\mathbf{P}\boldsymbol{X}_u] = 0$ for all $u \in [D]$.

Summand 1: $\|(80) - (81)\|_2$. Using the matrix Hoeffding inequality, we have with high probability, $\|(80) - (81)\|_2 \leqslant 16\sqrt{\log(d)\sum_{i=1}^{N} M_i^2}$, where $\left\|\frac{\eta Dy[i]}{N}\mathfrak{S}(-y[i]F_{\widehat{\boldsymbol{v}}}(\boldsymbol{X}[i]))\sum_{m=1}^{D}\left(D\sum_{r=1}^{D}\widehat{S}_{m,r}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_r[i]\rangle\right)^{p-1}\sum_{u=1}^{D}\widehat{S}_{m,u}^{(t)}\mathbf{P}\boldsymbol{X}_u[i]\right\|_2^2 \leqslant M_i^2$. Induction Hypothesis B.1, Lemma D.11 and $\|\mathbf{P}\boldsymbol{X}_u[i]\|_2 \leqslant \sigma^2 d\log(d)$ imply $M_i^2 \leqslant \frac{\eta^2\mathrm{poly}(D)}{N^2}$. We deduce that $\|(80) - (81)\|_2 \leqslant \eta\frac{\mathrm{poly}(D)}{\sqrt{N}}$.

Summand 2: $\|(82) - (83)\|_2$. The function $x \mapsto x^{p-1}$ is $(p-1)M^{p-2}$ Lipschitz on a bounded domain $[0, M]$. We apply this property and $(p-1)\max_{m\in[D]}\left(D\sum_{r=1}^{D}\widehat{S}_{m,r}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_r\rangle\right)^{p-2} \leqslant (\widehat{\alpha}^{(t)})^{p-2}\mathrm{poly}(D)$ to get:

$$
\begin{aligned}
&\|(82) - (83)\|_2 \\
\leqslant &\eta(\widehat{\alpha}^{(t)})^{p-2}\mathrm{poly}(D)\mathbb{E}\left[\sum_{m=1}^{D}\sum_{r=1}^{D}\widehat{S}_{m,r}^{(t)}|\langle\widehat{\boldsymbol{v}}^{(t)} - \mathring{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_r\rangle|\cdot\sum_{b=1}^{D}\widehat{S}_{m,b}^{(t)}\|\mathbf{P}\boldsymbol{X}_b\|_2\right] \\
\leqslant &\eta(\widehat{\alpha}^{(t)})^{p-2}\mathrm{poly}(D)\mathbb{E}\left[\sum_{m=1}^{D}\sum_{r=1}^{D}\widehat{S}_{m,r}^{(t)}\varepsilon_{\boldsymbol{v}}^{(t)}|\langle\boldsymbol{u}^{(t)}, \boldsymbol{X}_r\rangle|\cdot\sum_{b=1}^{D}\widehat{S}_{m,b}^{(t)}\|\mathbf{P}\boldsymbol{X}_b\|_2\right]. && (86)
\end{aligned}
$$

With high probability, $\|\boldsymbol{\xi}_b\|_2 \leqslant \sigma\sqrt{\log(d)} \leqslant \sqrt{\log(d)/d}$, $|\langle\boldsymbol{u}^{(t)}, \boldsymbol{\xi}_r\rangle| \leqslant \sqrt{\log(d)}\sigma$. We thus get:

$$
\|(82) - (83)\|_2 \leqslant \eta(\widehat{\alpha}^{(t)})^{p-2}\frac{\mathrm{poly}(D)}{\sqrt{d}}\varepsilon_{\boldsymbol{v}}^{(t)}. \tag{87}
$$

Summand 3: $\|(84)-(85)\|_2$**.** We have:

$$
\begin{aligned}
&\|(84)-(85)\|_2 \\
\leqslant \eta \Bigg\| \mathbb{E}\Bigg[& Dy \sum_{m=1}^{D} \Big(D \sum_{r=1}^{D} \widehat{S}_{m,r}^{(t)} \langle \mathring{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_r \rangle \Big)^{p-1} \sum_{b=1}^{D} \widehat{S}_{m,b}^{(t)} \mathbf{P} \boldsymbol{X}_u \cdot \\
& \Big[\mathfrak{S}\Big(-y \sum_{a=1}^{D} \big(D \sum_{a'=1}^{D} \widehat{S}_{a,a'}^{(t)} \langle \mathring{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_{a'} \rangle \big)^p \Big) - \mathfrak{S}\Big(-y \sum_{a=1}^{D} \big(D \sum_{a'=1}^{D} \widehat{S}_{a,a'}^{(t)} \langle \widehat{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_{a'} \rangle \big)^p \Big) \Big] \Bigg] \Bigg\|_2 .
\end{aligned}
\tag{88}
$$

We apply Lemma E.12 to bound the local change of the sigmoid in (88) which yields:

$$
\begin{aligned}
&\|(84)-(85)\|_2 \\
\leqslant & \Theta(\eta D) \mathbb{E}\Bigg[\sum_{m=1}^{D} \Big(D \sum_{r=1}^{D} \widehat{S}_{m,r}^{(t)} \langle \mathring{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_r \rangle \Big)^{p-1} \sum_{a=1}^{D} \Big| D \sum_{a'=1}^{D} \widehat{S}_{a,a'}^{(t)} \langle \mathring{\boldsymbol{v}}^{(t)} - \widehat{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_{a'} \rangle \Big| \cdot \sum_{b=1}^{D} \widehat{S}_{m,u}^{(t)} \|\mathbf{P}\boldsymbol{X}_b\|_2 \Bigg] \\
\leqslant & \Theta(\eta D^2) \mathbb{E}\Bigg[\sum_{m=1}^{D} \Big(D \sum_{r=1}^{D} \widehat{S}_{m,r}^{(t)} \langle \mathring{\boldsymbol{v}}^{(t)}, \boldsymbol{X}_r \rangle \Big)^{p-1} \sum_{a=1}^{D} \sum_{a'=1}^{D} \widehat{S}_{a,a'}^{(t)} \varepsilon_{\boldsymbol{v}}^{(t)} |\langle \boldsymbol{u}^{(t)}, \boldsymbol{X}_{a'} \rangle| \sum_{b=1}^{D} \widehat{S}_{m,u}^{(t)} \|\mathbf{P}\boldsymbol{X}_b\|_2 \Bigg].
\end{aligned}
\tag{89}
$$

We apply Induction Hypothesis B.1 to bound the softmax terms in (89). Besides, with high probability, we have $\|\boldsymbol{\xi}_b\|_2 \leqslant \sigma\sqrt{d\log(d)}$, $|\langle \boldsymbol{u}^{(t)}, \boldsymbol{\xi}_{a'} \rangle| \leqslant \sqrt{\log(d)}\sigma$. Thus, we have:

$$
\|(84)-(85)\|_2 \leqslant \eta \cdot \mathrm{poly}(D) (\widehat{\alpha}^{(t)})^{p-1} \varepsilon_{\boldsymbol{v}}^{(t)} \sigma \leqslant \eta \frac{\mathrm{poly}(D)}{\sqrt{d}} (\widehat{\alpha}^{(t)})^{p-1} \varepsilon_{\boldsymbol{v}}^{(t)}. \tag{90}
$$

We combine the bounds on the three summands to obtain the recursion of $\varepsilon_{\boldsymbol{v}}^{(t)}$. □

We now prove Lemma E.11 that gives the final bound on $\varepsilon_{\boldsymbol{v}}^{(t)}$ for $t \leqslant T$.

Lemma E.11. *For all $t \leqslant T$, $\varepsilon_{\boldsymbol{v}}^{(t)} \leqslant O(\frac{\mathrm{poly}(d)}{\sqrt{N}})$. By setting $N = \mathrm{poly}(d)$, $\varepsilon_{\boldsymbol{v}}^{(t)} \leqslant \frac{1}{\mathrm{poly}(d)}$.*

Proof of Lemma E.11. We bound $\varepsilon_{\boldsymbol{v}}^{(t)}$ in the following two regimes: $t \in [\mathscr{T}, \mathscr{T} + \check{\mathcal{T}}_0]$ and $t \in [\mathscr{T} + \check{\mathcal{T}}_0, T]$.

First phase: $t \in [\mathscr{T}, \mathscr{T} + \check{\mathcal{T}}_0]$**.** Unraveling Lemma E.10 for $t = \mathscr{T}, \ldots, \mathscr{T} + \check{\mathcal{T}}_0$ leads to:

$$
\begin{aligned}
\varepsilon_{\boldsymbol{v}}^{(t)} &\leqslant \Big[\varepsilon_{\boldsymbol{v}}^{(\mathscr{T})} + \eta\zeta\check{\mathcal{T}}_0 \Big] \prod_{\tau=\mathscr{T}}^{\mathscr{T}+\check{\mathcal{T}}_0} \Big(1 + \eta (\widehat{\alpha}^{(\tau)})^{p-1} \frac{\mathrm{poly}(D)}{\sqrt{d}} \Big) \\
&\leqslant [\varepsilon_{\boldsymbol{v}}^{(\mathscr{T})} + \eta\zeta\check{\mathcal{T}}_0] \prod_{\tau=0}^{\check{\mathcal{T}}_0-1} \Big(1 + \eta (\widehat{\alpha}^{(\tau)})^{p-1} \frac{\mathrm{poly}(D)}{\sqrt{d}} \Big).
\end{aligned}
\tag{91}
$$

Lemma E.20 provides the update of $\widehat{\alpha}^{(t)}$ during this time phase. We thus apply Lemma K.2 to bound the product term in (91).

$$
\prod_{\tau=0}^{\check{\mathcal{T}}_0-1} \Big(1 + \eta (\widehat{\alpha}^{(\tau)})^{p-1} \frac{\mathrm{poly}(D)}{\sqrt{d}} \Big) \leqslant \Big(1 + \frac{\Theta(1)}{C^{2(p-2)}\lambda_0^{p-2}} \Big)^{\frac{\eta \mathrm{poly}(D)}{\sqrt{d}}} \leqslant O(1). \tag{92}
$$

Plugging (92) in (91) yields a bound on $\varepsilon_{\boldsymbol{v}}^{(\mathscr{T}+\check{\mathcal{T}}_0)}$.

$$
\varepsilon_{\boldsymbol{v}}^{(\mathscr{T}+\check{\mathcal{T}}_0)} \leqslant O(1) \big(\varepsilon_{\boldsymbol{v}}^{(\mathscr{T})} + \eta\zeta\check{\mathcal{T}}_0 \big). \tag{93}
$$

Second phase: $t \in (\mathscr{T} + \check{\mathcal{T}}_0, T]$. Lemma E.23 shows that $\widehat{\alpha}^{(t)}$ gets updated until $t = \check{\mathcal{T}}_2 < T$. Therefore, we have $\varepsilon_{\boldsymbol{v}}^{(T)} = \varepsilon_{\boldsymbol{v}}^{(\mathscr{T}+\check{\mathcal{T}}_2)}$. Unraveling Lemma E.10 for $t = \mathscr{T} + \check{\mathcal{T}}_0, \ldots, \mathscr{T} + \check{\mathcal{T}}_2$ and using $\widehat{\alpha}^{(t)} \leqslant \frac{\mathrm{polylog}(d)}{C^2\lambda_0}$ for $t \leqslant T$ leads to:

$$\varepsilon_{\boldsymbol{v}}^{(\mathscr{T}+\check{\mathcal{T}}_2)} \leqslant \left(1 + \frac{\eta \mathrm{polylog}(d)}{\lambda_0^{p-1}} \frac{\mathrm{poly}(D)}{\sqrt{d}}\right)^{\check{\mathcal{T}}_2 - \check{\mathcal{T}}_0} \varepsilon_{\boldsymbol{v}}^{(\mathscr{T}+\check{\mathcal{T}}_0)} + \eta\zeta \frac{\left(1 + \frac{\eta \mathrm{polylog}(d)}{\lambda_0^{p-1}} \frac{\mathrm{poly}(D)}{\sqrt{d}}\right)^{\check{\mathcal{T}}_2 - \check{\mathcal{T}}_0} - 1}{\frac{\eta \mathrm{polylog}(d)}{\lambda_0^{p-1}} \frac{\mathrm{poly}(D)}{\sqrt{d}}}. \tag{94}$$

Since $\check{\mathcal{T}}_2 - \check{\mathcal{T}}_0 \leqslant \frac{\lambda_0^p}{\eta e^{\mathrm{polyloglog}(d)}}$, we have $\left(1 + \frac{\eta \mathrm{polylog}(d)}{\lambda_0^{p-1}} \frac{\mathrm{poly}(D)}{\sqrt{d}}\right)^{\check{\mathcal{T}}_2 - \check{\mathcal{T}}_0} \leqslant O(1)$. Simplifying (94) yields:

$$\varepsilon_{\boldsymbol{v}}^{(\mathscr{T}+\check{\mathcal{T}}_2)} \leqslant O\left(\frac{\mathrm{poly}(d)}{\sqrt{N}}\right). \tag{95}$$

(95) implies that we need $N = \mathrm{poly}(d)$ samples to have $\varepsilon_{\boldsymbol{v}}^{(T)} \leqslant 1/\mathrm{poly}(d)$. □

E.2.3 Auxiliary lemmas

In this section, we prove the Lipschitzness of the function appearing in the proof of Lemma E.10.

Lemma E.12. *Let $\psi\colon \mathbb{R}^d \to \mathbb{R}$ defined as $\psi(\boldsymbol{x}) := \mathfrak{S}(-\sum_{m=1}^D x_m^p)$ and $p \geqslant 3$ be an odd integer. Assume that $x_m^p \geqslant -1/D$. Then, ψ is p-Lipschitz i.e. for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $|\psi(\boldsymbol{x}) - \psi(\boldsymbol{y})| \leqslant p\|\boldsymbol{x} - \boldsymbol{y}\|_1$.*

Proof of Lemma E.12. Let $l \in [D]$. The derivative of ψ with respect to a variable x_l is:

$$\left|\frac{\partial \psi}{\partial x_l}\right| = p\frac{x_l^{p-1}\exp(\sum_{m=1}^D x_m^p)}{(1+\exp(\sum_{m=1}^D x_m^p))^2} \leqslant \frac{p x_l^{p-1}}{1+\exp(\sum_{m=1}^D x_m^p)}. \tag{96}$$

(96) implies a bound on $\|\nabla\psi(\boldsymbol{x})\|_1$. Indeed, since $\sum_{m=1}^D x_m^p \geqslant -1$, we have:

$$\|\nabla\psi(\boldsymbol{x})\|_1 \leqslant p\frac{\sum_{m=1}^D (yx_m)^{p-1}}{1+\exp(\sum_{m=1}^D (yx_m)^p)} \leqslant p. \tag{97}$$

(97) shows that ψ is p-Lipschitz. □

E.2.4 $\Delta_{\boldsymbol{A}}^{(t)}$ stays small during the learning process

Here, we bound the gap in attention coefficients between the realistic and semi-idealized cases.

Lemma E.13. *Assume that we run GD on the empirical risk (E) for T iterations with parameters set as in Parametrization 3.1 and the number of samples is $N = \mathrm{poly}(d)$. Then, the attention matrix in the realistic case is very close to the semi-idealized one i.e. for $t \in [\mathscr{T}, T]$,*

$$\Delta_{\boldsymbol{A}}^{(t)} := \max_{i \neq j} |\widehat{A}_{i,j}^{(t)} - A_{i,j}^{(t)}| \leqslant \frac{1}{\mathrm{poly}(d)}.$$

We now detail the steps to prove Lemma E.13. We first provide the recursion that $\Delta_{\boldsymbol{A}}^{(t)}$ satisfies.

Lemma E.14. *Assume that we run GD on the empirical risk (E) for T iterations with parameters set as in Parametrization 3.1. Then, the discrepancy $\Delta_{\boldsymbol{A}}$ satisfies for $t \in (\mathscr{T}, T]$,*

$$\Delta_{\boldsymbol{A}}^{(t+1)} \leqslant \left(1 + \eta R^{(t)}\right)\Delta_{\boldsymbol{A}}^{(t)} + \eta \mathrm{poly}(D)\sigma\varepsilon_{\boldsymbol{v}}^{(t)} + \eta\zeta,$$

where $\Delta_{\boldsymbol{A}}^{(\mathscr{T})} \leqslant |\widehat{A}_{i,j}^{(\mathscr{T})}|$, $R^{(t)} = O(C)(G^{(t)})^{p-1}\check{\Gamma}^{(t)}(\widehat{\alpha}^{(t)})^p$ and $\zeta = \frac{\mathrm{poly}(D)}{\sqrt{N}}$.

Proof of Lemma E.14. In this proof, we maintain the hypothesis that $\Delta_{\boldsymbol{A}}^{(t)}$ is small. We will eventually prove this statement in Lemma E.15. Let $a, b \in [D]$ such that $a \neq b$. Using GD, $\widehat{A}_{a,b}^{(t+1)} - \check{A}_{a,b}^{(t+1-\mathscr{T})}$ satisfies:

$$\begin{aligned}
&\big|\widehat{A}_{a,b}^{(t+1)} - \check{A}_{a,b}^{(t+1-\mathscr{T})}\big| \leqslant \big|\widehat{A}_{a,b}^{(t)} - \check{A}_{a,b}^{(t-\mathscr{T})}\big| \\
&+ \eta\Bigg|\frac{D}{N}\sum_{i=1}^{N} y[i]\mathfrak{S}\big(-y[i]F_{\widehat{\boldsymbol{v}},\widehat{\boldsymbol{A}}}(\boldsymbol{X}[i])\big)\Big(D\sum_{c=1}^{D}\widehat{S}_{a,c}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c[i]\rangle\Big)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b[i]-\boldsymbol{X}_r[i]\rangle \qquad (98)\\
&\quad -\mathbb{E}\Big[Dy\mathfrak{S}(-yF_{\widehat{\boldsymbol{v}},\widehat{\boldsymbol{A}}}(\boldsymbol{X}))\Big(D\sum_{c=1}^{D}\widehat{S}_{a,c}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\Big)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle\Big]\Bigg| \qquad (99)\\
&+ \eta\Bigg|\mathbb{E}\Big[Dy\mathfrak{S}(-yF_{\widehat{\boldsymbol{v}},\widehat{\boldsymbol{A}}}(\boldsymbol{X}))\Big(D\sum_{c=1}^{D}\widehat{S}_{a,c}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\Big)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle\Big] \qquad (100)\\
&\quad -\mathbb{E}\Big[Dy\mathfrak{S}(-yF_{\check{\boldsymbol{v}},\widehat{\boldsymbol{A}}}(\boldsymbol{X}))\Big(D\sum_{c=1}^{D}\widehat{S}_{a,c}^{(t)}\langle\check{\boldsymbol{v}}^{(t-\mathscr{T})},\boldsymbol{X}_c\rangle\Big)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}\langle\check{\boldsymbol{v}}^{(t-\mathscr{T})},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle\Big]\Bigg| \qquad (101)\\
&+ D\eta\Bigg|\mathbb{E}\Big[y\mathfrak{S}(-yF_{\check{\boldsymbol{v}},\widehat{\boldsymbol{A}}}(\boldsymbol{X}))\Big(D\sum_{c=1}^{D}\widehat{S}_{a,c}^{(t)}\langle\check{\boldsymbol{v}}^{(t-\mathscr{T})},\boldsymbol{X}_c\rangle\Big)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle\Big] \qquad (102)\\
&- \mathbb{E}\Big[y\mathfrak{S}(-yF_{\check{\boldsymbol{v}},\check{\boldsymbol{A}}}(\boldsymbol{X}))\Big(D\sum_{c=1}^{D}\check{S}_{a,c}^{(t-\mathscr{T})}\langle\check{\boldsymbol{v}}^{(t-\mathscr{T})},\boldsymbol{X}_c\rangle\Big)^{p-1}\check{S}_{a,b}^{(t-\mathscr{T})}\sum_{r\neq b}\check{S}_{a,r}^{(t-\mathscr{T})}\langle\check{\boldsymbol{v}}^{(t-\mathscr{T})},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle\Big]\Bigg|. \qquad (103)
\end{aligned}$$

We now bound the three summands above.

Summand 1: $|(98)-(99)|$. We apply the Hoeffding inequality. With high probability, we have: $|(98)-(99)| \leqslant \eta\frac{\mathrm{poly}(D)}{\sqrt{N}}$.

Summand 2: $|(100)-(101)|$. Since $\varepsilon_{\boldsymbol{v}}^{(t)}$ is small (Lemma E.11), we can show that:

$$|(100)-(101)| \leqslant \eta D\mathbb{E}\Bigg[\Big(D\sum_{c=1}^{D}\widehat{S}_{a,c}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\Big)^{p-1}\widehat{S}_{a,b}^{(t)}\cdot\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}|\langle\boldsymbol{u}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle|\Bigg]\varepsilon_{\boldsymbol{v}}^{(t)}. \qquad (104)$$

With high probability, we have $\langle\boldsymbol{u}^{(t)},\boldsymbol{\xi}_b\rangle \leqslant \sigma\sqrt{\log(d)}$. Using this fact along with $D\Big|\Big(D\sum_{c=1}^{D}\widehat{S}_{a,c}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\Big)^{p-1}\widehat{S}_{a,b}^{(t)}\cdot\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}\Big| \leqslant \mathrm{poly}(D)$, we further bound (104) as:

$$|(100)-(101)| \leqslant \eta\mathrm{poly}(D)\varepsilon_{\boldsymbol{v}}^{(t)}\sigma\sqrt{\log(d)}. \qquad (105)$$

Summand 3: $|(102)-(103)|$**.** We have the following decomposition.

$$|(102)-(103)|$$

$$\leqslant \eta\Bigg|\mathbb{E}\Bigg[Dy\mathfrak{S}(-yF_{\check{\boldsymbol{v}},\hat{\boldsymbol{A}}}(\boldsymbol{X}))\Bigg(D\sum_{c=1}^{D}\widehat{S}_{a,c}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\Bigg)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle\Bigg] \tag{106}$$

$$-\mathbb{E}\Bigg[Dy\mathfrak{S}(-yF_{\check{\boldsymbol{v}},\check{\boldsymbol{A}}}(\boldsymbol{X}))\Bigg(D\sum_{c=1}^{D}\widehat{S}_{a,c}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\Bigg)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle\Bigg] \tag{107}$$

$$+\eta\Bigg|\mathbb{E}\Bigg[Dy\mathfrak{S}(-yF_{\check{\boldsymbol{v}},\check{\boldsymbol{A}}}(\boldsymbol{X}))\Bigg(D\sum_{c=1}^{D}\widehat{S}_{a,c}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\Bigg)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle\Bigg] \tag{108}$$

$$-\mathbb{E}\Bigg[Dy\mathfrak{S}(-yF_{\check{\boldsymbol{v}},\check{\boldsymbol{A}}}(\boldsymbol{X}))\Bigg(D\sum_{c=1}^{D}\check{S}_{a,c}^{(t-\mathscr{T})}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\Bigg)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle\Bigg]\Bigg| \tag{109}$$

$$+\eta\Bigg|\mathbb{E}\Bigg[Dy\mathfrak{S}(-yF_{\check{\boldsymbol{v}},\check{\boldsymbol{A}}}(\boldsymbol{X}))\Bigg(D\sum_{c=1}^{D}\check{S}_{a,c}^{(t-\mathscr{T})}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\Bigg)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle\Bigg] \tag{110}$$

$$-\mathbb{E}\Bigg[Dy\mathfrak{S}(-yF_{\check{\boldsymbol{v}},\check{\boldsymbol{A}}}(\boldsymbol{X}))\Bigg(D\sum_{c=1}^{D}S_{a,c}^{(t-\mathscr{T})}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\Bigg)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}\check{S}_{a,r}^{(t-\mathscr{T})}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle\Bigg]\Bigg| \tag{111}$$

$$+\eta\Bigg|\mathbb{E}\Bigg[Dy\mathfrak{S}(-yF_{\check{\boldsymbol{v}},\check{\boldsymbol{A}}}(\boldsymbol{X}))\Bigg(D\sum_{c=1}^{D}\check{S}_{a,c}^{(t-\mathscr{T})}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\Bigg)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}\check{S}_{a,r}^{(t-\mathscr{T})}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle\Bigg]\Bigg| \tag{112}$$

$$-\mathbb{E}\Bigg[Dy\mathfrak{S}(-yF_{\check{\boldsymbol{v}},\check{\boldsymbol{A}}}(\boldsymbol{X}))\Bigg(D\sum_{c=1}^{D}\check{S}_{a,c}^{(t-\mathscr{T})}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\Bigg)^{p-1}\check{S}_{a,b}^{(t-\mathscr{T})}\sum_{r\neq b}\check{S}_{a,r}^{(t-\mathscr{T})}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle\Bigg]\Bigg|. \tag{113}$$

We need to distinguish two sub-cases: $a,b\in\mathcal{S}_\ell$ and $a\in\mathcal{S}_\ell$, $b\in\mathcal{S}_m$ with $\ell\neq m$.

Subcase 1: $a,b\in\mathcal{S}_\ell$**.** The proof of Lemma G.1 highlights that when $a,b\in\mathcal{S}_\ell$ the event with largest gradient is event a: "$\ell=\ell(\pi(\boldsymbol{X}))$" which happens with probability $1/L$. Therefore, to simplify the calculations, we will only take into account this event. We first bound $|(106)-(107)|$. We successively apply Lemma E.12 (Lipschitzness of sigmoid) and Lemma E.17 (Lipschitzness of softmax) and get:

$$|(106)-(107)|$$

$$\leqslant\frac{\Theta(\eta D)}{L}\mathbb{E}\Bigg[\Bigg(D\sum_{c=1}^{D}\widehat{S}_{a,c}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\Bigg)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{c=1}^{D}\sum_{c'=1}^{D}|\check{S}_{c,c'}^{(t-\mathscr{T})}-\widehat{S}_{c,c'}^{(t)}|\cdot|\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_{c'}\rangle|\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}|\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle|\,\Bigg|\,\mathbf{a}\Bigg]$$

$$\leqslant\frac{\Theta(\eta D)}{L}\mathbb{E}\Bigg[\Bigg(D\sum_{c=1}^{D}\widehat{S}_{a,c}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\Bigg)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{c=1}^{D}\sum_{c'=1}^{D}\check{S}_{c,c'}^{(t-\mathscr{T})}|e^{2\Delta_{\boldsymbol{A}}^{(t)}}-1||\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_{c'}\rangle|\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}|\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle|\,\Bigg|\,\mathbf{a}\Bigg]$$

$$\leqslant\frac{\Theta(\eta D)}{L}\mathbb{E}\Bigg[\Bigg(D\sum_{c=1}^{D}\widehat{S}_{a,c}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\Bigg)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{c=1}^{D}\sum_{c'=1}^{D}\check{S}_{c,c'}^{(t-\mathscr{T})}|\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_{c'}\rangle|\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}|\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle|\,\Bigg|\,\mathbf{a}\Bigg]\Delta_{\boldsymbol{A}}^{(t)} \tag{114}$$

$$\leqslant\frac{\Theta(\eta\widehat{\alpha}^{(t)}\lambda_0D^2)}{L}\mathbb{E}\Bigg[\Bigg(D\sum_{c=1}^{D}\widehat{S}_{a,c}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\Bigg)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}|\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle|\,\Bigg|\,\mathbf{a}\Bigg]\Delta_{\boldsymbol{A}}^{(t)}. \tag{115}$$

where we used $e^{\Delta_{\boldsymbol{A}}^{(t)}}-1\leqslant 2\Delta_{\boldsymbol{A}}^{(t)}$ in (114) and Induction Hypothesis B.1 in (115). Using Lipschitz inequalities, we can further expand (115) as a function of the coefficients from $\check{S}^{(t)}$ and $\Delta_{\boldsymbol{A}}^{(t)}$.

However, $\Delta_{\boldsymbol{A}}^{(t)}$ is small and we only want terms of order 1 in $\Delta_{\boldsymbol{A}}^{(t)}$ in (115). Therefore, the only term of order 1 that remains is:

$$|(106)-(107)|$$
$$\leqslant \frac{\Theta(\eta\widehat{\alpha}^{(t)}\lambda_0 D^2)}{L}\mathbb{E}\Bigg[\left(D\sum_{c=1}^{D}\check{S}_{a,c}^{(t-\mathscr{T})}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\right)^{p-1}\check{S}_{a,b}^{(t-\mathscr{T})}\sum_{r\neq b}\check{S}_{a,r}^{(t-\mathscr{T})}|\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle|\,\Bigg|\,\mathbf{a}\Bigg]\Delta_{\boldsymbol{A}}^{(t)}. \tag{116}$$

Bounding the expectation in (116) as in the proof of Lemma G.1 yields $|(106)-(107)|\leqslant \eta R^{(t)}\Delta_{\boldsymbol{A}}^{(t)}$. We now bound $|(108)-(109)|$. We therefore apply (Lemma E.16) and get:

$$|(108)-(109)| \tag{117}$$
$$\leqslant\frac{\Theta(\eta D)}{L}\mathbb{E}\Bigg[\left(D\sum_{c=1}^{D}\check{S}_{a,c}^{(t-\mathscr{T})}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\right)^{p-1}|e^{2(p-1)\Delta_{\boldsymbol{A}}^{(t)}}-1|\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle\,\Bigg|\,\mathbf{a}\Bigg] \tag{118}$$
$$\leqslant\frac{\Theta(\eta D)}{L}\mathbb{E}\Bigg[\left(D\sum_{c=1}^{D}\check{S}_{a,c}^{(t-\mathscr{T})}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\right)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}\widehat{S}_{a,r}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle\,\Bigg|\,\mathbf{a}\Bigg]\Delta_{\boldsymbol{A}}^{(t)}. \tag{119}$$

where we used $e^{2(p-1)\Delta_{\boldsymbol{A}}^{(t)}}-1\leqslant 4(p-1)\Delta_{\boldsymbol{A}}^{(t)}$ in (119). We can further expand (119), keep the terms of first order in $\Delta_{\boldsymbol{A}}$ and get $|(110)-(111)|\leqslant\eta R^{(t)}\Delta_{\boldsymbol{A}}^{(t)}$.

We now bound $|(110)-(111)|$. Using the Lipschitz property of the softmax (Lemma E.17), we have:

$$|(110)-(111)|$$
$$\leqslant\frac{\eta D}{L}\mathbb{E}\Bigg[\left(D\sum_{c=1}^{D}\check{S}_{a,c}^{(t-\mathscr{T})}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\right)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}|e^{2\Delta_{\boldsymbol{A}}^{(t)}}-1|\cdot\check{S}_{a,r}^{(t-\mathscr{T})}\cdot|\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle|\Bigg|\mathbf{a}\Bigg]$$
$$\leqslant\frac{\Theta(\eta D)}{L}\mathbb{E}\Bigg[\left(D\sum_{c=1}^{D}\check{S}_{a,c}^{(t-\mathscr{T})}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\right)^{p-1}\widehat{S}_{a,b}^{(t)}\sum_{r\neq b}\check{S}_{a,r}^{(t-\mathscr{T})}|\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle|\Bigg|\mathbf{a}\Bigg]\Delta_{\boldsymbol{A}}^{(t)}, \tag{120}$$

where we used $|e^{2\Delta_{\boldsymbol{A}}^{(t)}}-1|\leqslant 4\Delta_{\boldsymbol{A}}^{(t)}$ in (120). Using the same arguments as above, we obtain $|(110)-(111)|\leqslant\eta R^{(t)}\Delta_{\boldsymbol{A}}^{(t)}$.

The bound on $|(112)-(113)|$ can be derived as above. We again use the Lipschitz property of softmax (Lemma E.17) which leads to

$$|(112)-(113)|\leqslant\frac{\Theta(\eta D)}{L}\mathbb{E}\Bigg[\left(D\sum_{c=1}^{D}\check{S}_{a,c}^{(t-\mathscr{T})}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_c\rangle\right)^{p-1}\check{S}_{a,b}^{(t-\mathscr{T})}\sum_{r\neq b}\check{S}_{a,r}^{(t-\mathscr{T})}|\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_b-\boldsymbol{X}_r\rangle|\Bigg|\mathbf{a}\Bigg]\Delta_{\boldsymbol{A}}^{(t)}$$
$$\leqslant\eta R^{(t)}\Delta_{\boldsymbol{A}}^{(t)}. \tag{121}$$

Subcase 2: $a\in\mathcal{S}_\ell$ and $b\in\mathcal{S}_m$ with $\ell\neq m$. The proof is analogous to the Subcase 1. We only take into account event a: "$\ell=\ell(\pi(\boldsymbol{X}))$ and $\delta_j=0$" and **Event e**: "$\ell,m\neq\ell(\pi(\boldsymbol{X}))$ and $\delta_j=0$ and $\delta_s=0$ and show that $|(102)-(103)|\leqslant\eta\tilde{R}^{(t)}\Delta_{\boldsymbol{A}}^{(t)}$ where $\tilde{R}^{(t)}:=O(\widehat{\alpha}^{(t)})\left(D\widehat{\alpha}^{(t)}(\Lambda^{(t)}+(C-1)\Gamma^{(t)})\right)^{p-1}\Xi^{(t)}\frac{\lambda_0}{D}+O(\widehat{\alpha}^{(t)})\left(D\widehat{\alpha}^{(t)}qD\log(d)\Xi^{(t)}\right)^{p-1}\Xi^{(t)}qD\log(d)$.

Putting all the pieces together. Given the value of the parameters, we know that $\tilde{R}^{(t)}\leqslant R^{(t)}$ for all $t\in[T]$. Therefore, Summand 3 is bounded as:

$$|(102)-(103)|\leqslant\eta R^{(t)}\Delta_{\boldsymbol{A}}^{(t)}. \tag{122}$$

Conclusion. Plugging the bounds on Summands 1, 2 and 3 in the original decomposition of $\left|\widehat{A}_{a,b}^{(t+1)}-\check{A}_{a,b}^{(t+1-\mathscr{T})}\right|$ yields the bound on $\Delta_{\boldsymbol{A}}^{(t)}$. The second part of the lemma is obtained using Lemma E.15.

□

Lemma E.15. *Let $N = \mathrm{poly}(d)$. Then, for all $t \leqslant T$, $\Delta_{\boldsymbol{A}}^{(t)} \leqslant \frac{1}{\mathrm{poly}(d)}$.*

Proof of Lemma E.15. Let $\mathcal{E}_{\boldsymbol{v}} > 0$ such that $\varepsilon_{\boldsymbol{v}}^{(t)} \leqslant \mathcal{E}_{\boldsymbol{v}}$ for $t \in [T]$ – we proved the existence of $\mathcal{E}_{\boldsymbol{v}}$ in Lemma E.11. We bound $\Delta_{\boldsymbol{A}}^{(t)}$ when $t \in [\mathscr{T}, \mathscr{T} + \check{\mathcal{T}}_0]$ and $t \in [\mathscr{T} + \check{\mathcal{T}}_0, T]$.

First phase: $t \in [\mathscr{T}, \mathscr{T} + \check{\mathcal{T}}_0]$**.** Unraveling Lemma E.14 for $t = \mathscr{T}, \dots, \mathscr{T} + \check{\mathcal{T}}_0$ leads to:

$$\begin{aligned}
&\Delta_{\boldsymbol{A}}^{(\mathscr{T}+\check{\mathcal{T}}_0)} \\
&\leqslant \Big(\Delta_{\boldsymbol{A}}^{(\mathscr{T})} + \eta\check{\mathcal{T}}_0\big(\mathcal{E}_{\boldsymbol{v}}\mathrm{poly}(D)\sigma + \zeta\big)\Big)\prod_{\tau=\mathscr{T}}^{\mathscr{T}+\check{\mathcal{T}}_0-1}\Big(1 + \frac{\eta D(\widehat{\alpha}^{(\tau)})^p}{L}O\big(D(\check{\Lambda}^{(t)} + (C-1)\check{\Gamma}^{(t)})\big)^{p-1}\check{\Gamma}^{(t)}\Big).
\end{aligned} \tag{123}$$

We now apply Induction Hypothesis B.1 to simplify (123) and get:

$$\Delta_{\boldsymbol{A}}^{(\mathscr{T}+\check{\mathcal{T}}_0)} \leqslant \Big(\Delta_{\boldsymbol{A}}^{(\mathscr{T})} + \eta\check{\mathcal{T}}_0\big(\mathcal{E}_{\boldsymbol{v}}\mathrm{poly}(D)\sigma + \zeta\big)\Big)\prod_{\tau=\mathscr{T}}^{\mathscr{T}+\check{\mathcal{T}}_0-1}\Big(1 + \frac{O(\eta\lambda_0^p)}{L}(\widehat{\alpha}^{(t)})^p\Big). \tag{124}$$

We then apply Lemma K.2 to bound the product term in (124). We obtain:

$$\begin{aligned}
\Delta_{\boldsymbol{A}}^{(\mathscr{T}+\mathcal{T}_0)} &\leqslant \Big(\Delta_{\boldsymbol{A}}^{(\mathscr{T})} + \eta\check{\mathcal{T}}_0\big(\mathcal{E}_{\boldsymbol{v}}\mathrm{poly}(D)\sigma + \zeta\big)\Big)\Big(1 + \frac{\Theta(1)}{C^{2(p-2)}\lambda_0^{p-2}}\Big)^{\frac{\eta\lambda_0^{2p}}{L}} \\
&\leqslant O\Big(\Delta_{\boldsymbol{A}}^{(\mathscr{T})} + \eta\check{\mathcal{T}}_0\big(\mathcal{E}_{\boldsymbol{v}}\mathrm{poly}(D)\sigma + \zeta\big)\Big).
\end{aligned} \tag{125}$$

Second phase: $t \in [\mathscr{T} + \check{\mathcal{T}}_0, T]$**.** Unraveling Lemma E.14 for $t = \mathscr{T} + \check{\mathcal{T}}_0, \dots, \mathscr{T} + \check{\mathcal{T}}_2$ and and using $\widehat{\alpha}^{(t)} \leqslant \frac{\mathrm{polylog}(d)}{C^2\lambda_0}$ for $t \leqslant T$ leads to:

$$\Delta_{\boldsymbol{A}}^{(\mathscr{T}+\check{\mathcal{T}}_2)} \leqslant \Big(1 + \frac{\eta\mathrm{polylog}(d)}{L}\Big)^{\check{\mathcal{T}}_2-\check{\mathcal{T}}_0}\Big(\Delta_{\boldsymbol{A}}^{(\mathscr{T}+\check{\mathcal{T}}_0)} + L\cdot\mathrm{polylog}(d)\big(\mathcal{E}_v\mathrm{poly}(D)\sigma + \zeta\big)\Big). \tag{126}$$

Using $\check{\mathcal{T}}_2 - \check{\mathcal{T}}_0 \leqslant \frac{\lambda_0^p}{\eta e^{\mathrm{polyloglog}(d)}}$, we have $\Big(1 + \frac{\eta\mathrm{polylog}(d)}{L}\Big)^{\check{\mathcal{T}}_2-\check{\mathcal{T}}_0} \leqslant O(1)$. We thus bound (125) as:

$$\Delta_{\boldsymbol{A}}^{(\mathscr{T}+\check{\mathcal{T}}_2)} \leqslant O(\Delta_{\boldsymbol{A}}^{(\mathscr{T})}) + \mathrm{polylog}(d)\Big[\mathcal{E}_{\boldsymbol{v}}\sigma + \frac{1}{\sqrt{N}}\Big]\Big[\frac{1}{C(\widehat{\alpha}^{(\mathscr{T})})^{p-1}e^{\mathrm{polyloglog}(d)}} + L\Big] \leqslant O\Big(\frac{\nu}{\sqrt{N}}\Big). \tag{127}$$

We deduce that setting $N = \mathrm{poly}(d)$ yields $\Delta_{\boldsymbol{A}}^{(T)} = \Delta_{\boldsymbol{A}}^{(\mathscr{T}+\check{\mathcal{T}}_2)} \leqslant 1/\mathrm{poly}(d)$. □

E.2.5 Auxiliary lemmas

Lemma E.16. *Let $\psi\colon \mathbb{R}^D \to \mathbb{R}$ defined as $\psi(\boldsymbol{x}) := (D\sum_{m=1}^D S_m\langle\boldsymbol{v}, \boldsymbol{X}_m\rangle)^{p-1}$ where $S_m = (\mathrm{softmax}(x_1, \dots, x_D))_m$, $\{\boldsymbol{X}_m\}_{m=1}^D$ and $\boldsymbol{v}$ are fixed vectors and $p \geqslant 3$ is an odd integer. Then, we have:*

$$|\psi(\boldsymbol{x}) - \psi(\boldsymbol{y})| \leqslant \psi(\boldsymbol{x})\big|e^{2(p-1)\max_{c\in[D]}|x_c - y_c|} - 1\big|.$$

Proof of Lemma E.16. Let $S_c = (\mathrm{softmax}(x_1, \dots, x_D))_c$, $\widehat{S}_c = (\mathrm{softmax}(y_1, \dots, y_D))_c$. We have:

$$\begin{aligned}
|\psi(\boldsymbol{x}) - \psi(\boldsymbol{y})| &\leqslant \psi(\boldsymbol{x})\left|\left[\frac{\sum_{c=1}^D S_c\langle\boldsymbol{v}, \boldsymbol{X}_c\rangle}{\sum_{c=1}^D \widehat{S}_c\langle\boldsymbol{v}, \boldsymbol{X}_c\rangle}\right]^{p-1} - 1\right| && (128) \\
&\leqslant \psi(\boldsymbol{x})\left|\left[\frac{\sum_{c=1}^D e^{x_c}\langle\boldsymbol{v}, \boldsymbol{X}_c\rangle}{\sum_{c=1}^D e^{y_c}\langle\boldsymbol{v}, \boldsymbol{X}_c\rangle}\cdot\frac{\sum_{r=1}^D e^{y_r}}{\sum_{r=1}^D e^{x_r}}\right]^{p-1} - 1\right|, && (129)
\end{aligned}$$

We finally apply the generalized mediant inequality in (129) and get:

$$|\psi(\boldsymbol{x}) - \psi(\boldsymbol{y})| \leqslant \psi(\boldsymbol{x})\Big|\Big(\max_{c\in[D]} e^{|x_c - y_c|} \cdot \max_{r\in[D]} e^{|x_r - y_r|}\Big)^{p-1} - 1\Big|$$
$$\leqslant \psi(\boldsymbol{x})\big|e^{2(p-1)\max_{c\in[D]}|x_c - y_c|} - 1\big|.$$

□

Lemma E.17 (Lipschitzness of Softmax)**.** *Let $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^D$. For all $i \in [D]$, we have:*

$$|\mathrm{softmax}(\boldsymbol{a})_i - \mathrm{softmax}(\boldsymbol{b})_i| \leqslant \big|e^{2\max_{k\in[D]}|b_k - a_k|} - 1\big| \cdot \mathrm{softmax}(\boldsymbol{a})_i.$$

Proof of Lemma E.17. Let $i \in [D]$. The difference of softmax is bounded as:

$$|\mathrm{softmax}(\boldsymbol{a})_i - \mathrm{softmax}(\boldsymbol{b})_i| \leqslant \frac{e^{a_i}}{\sum_{j=1}^D e^{a_j}} \cdot \left|\frac{e^{b_i}}{e^{a_i}} \frac{\sum_{j=1}^D e^{a_j}}{\sum_{j=1}^D e^{b_j}} - 1\right|$$
$$\leqslant \frac{e^{a_i}}{\sum_{j=1}^D e^{a_j}} \cdot \Big|e^{b_i - a_i} \max_{k\in[D]} e^{a_k - b_k} - 1\Big|, \tag{130}$$

where we used the mediant inequality in the last inequality of (130). Since the exponential function is non-decreasing, we deduce:

$$|\mathrm{softmax}(\boldsymbol{a})_i - \mathrm{softmax}(\boldsymbol{b})_i| \leqslant \big|e^{2\max_{k\in[D]}|b_k - a_k|} - 1\big| \cdot \mathrm{softmax}(\boldsymbol{a})_i. \tag{131}$$

□

E.2.6 Dynamics of $\widehat{\boldsymbol{v}}^{(t)}$

Lastly, since $\Delta_{\boldsymbol{A}}^{(t)}$ remains small and $\widehat{\boldsymbol{v}}^{(t)}$ mainly lies in $\mathrm{span}(\boldsymbol{w}^*)$, we show that $\widehat{\alpha}^{(t)}$ satisfies the same updates as the ideal $\alpha^{(t)}$ (up to some constant factors).

Lemma E.18. *Assume that we run GD on the empirical risk (E) for T iterations with parameters set as in Parametrization 3.1 and the number of samples is $N = \mathrm{poly}(d)$. Then, there exist times $\mathscr{T}, \widehat{\mathcal{T}}_0, \widehat{\mathcal{T}}_1 > 0$ such that*

1. *Analog of Event I (Lemma D.2): $\widehat{\alpha}^{(t+1)} = \widehat{\alpha}^{(t)} + \Theta(\eta)(\widehat{G}^{(t)})^p(\widehat{\alpha}^{(t)})^{p-1}$ for $t \in [\mathscr{T}, \mathscr{T} + \widehat{\mathcal{T}}_0]$.*
2. *Analog of Event III (Lemma D.11): $\widehat{\alpha}^{(t+1)} = \widehat{\alpha}^{(t)} + \Theta(\eta)(\widehat{G}^{(t)})^p(\widehat{\alpha}^{(t)})^{p-1}$ for $t \in [\mathscr{T} + \widehat{\mathcal{T}}_1, T]$.*

Consequently, $\widehat{\alpha}^{(t)}$ is non-decreasing and eventually, $\widehat{\alpha}^{(T)} = \alpha^{(T)} = \frac{\mathrm{polylog}(d)}{C^2\lambda_0}$.

These three lemmas imply that *at time T*, the realistic iterates are very close to the ideal ones. Therefore, they incur nearby test loss and thus the realistic model generalizes. We now proceed to the proof of

In order to analyze the dynamics of $\widehat{\boldsymbol{v}}^{(t)}$, we first show that the gradient (with respect to $\widehat{\boldsymbol{v}}$) in the realistic learning process is very close to the one in the semi-idealized one.

Lemma E.19. *Let $t \in [\mathscr{T}, T]$. With high probability, we have*

$$\|\nabla_{\widehat{\boldsymbol{v}}}\widehat{\mathcal{L}}(\widehat{\boldsymbol{v}}^{(t)}, \widehat{\boldsymbol{A}}^{(t)}) - \nabla_{\breve{\boldsymbol{v}}}\breve{\mathcal{L}}(\breve{\boldsymbol{v}}^{(t-\mathscr{T})}, \breve{\boldsymbol{A}}^{(t-\mathscr{T})})\|_2 \leqslant \mathrm{poly}(D)\Big(\frac{1}{\sqrt{N}} + \varepsilon_{\boldsymbol{v}}^{(t)} + \Delta_{\boldsymbol{A}}^{(t)}\Big). \tag{132}$$

By choosing $N = \mathrm{poly}(d)$, we have $\|\nabla_{\widehat{\boldsymbol{v}}}\widehat{\mathcal{L}}(\widehat{\boldsymbol{v}}^{(t)}, \widehat{\boldsymbol{A}}^{(t)}) - \nabla_{\breve{\boldsymbol{v}}}\breve{\mathcal{L}}(\breve{\boldsymbol{v}}^{(t-\mathscr{T})}, \breve{\boldsymbol{A}}^{(t-\mathscr{T})})\|_2 \leqslant 1/\mathrm{poly}(d)$.

Proof of Lemma E.19. We have:

$$\|\nabla_{\widehat{\boldsymbol{v}}}\widehat{\mathcal{L}}(\widehat{\boldsymbol{v}},\widehat{\boldsymbol{A}})-\nabla_{\check{\boldsymbol{v}}}\check{\mathcal{L}}(\check{\boldsymbol{v}},\check{\boldsymbol{A}})\|_2 \tag{133}$$

$$=\Big|\frac{D}{N}\sum_{i=1}^{N}y[i]\mathfrak{S}\big(-y[i]F_{\widehat{\boldsymbol{A}}^{(t)},\widehat{\boldsymbol{v}}^{(t)}}(\boldsymbol{X}[i])\big)\sum_{j=1}^{D}\sigma'\Big(\sum_{k=1}^{D}\widehat{S}_{j,k}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)},\boldsymbol{X}_k[i]\rangle\Big)\sum_{r=1}^{D}\widehat{S}_{j,r}^{(t)}\langle\boldsymbol{X}_r[i],\boldsymbol{w}^*\rangle \tag{134}$$

$$-D\mathbb{E}\Big[y\mathfrak{S}\big(-yF_{\widehat{\boldsymbol{A}}^{(t)},\widehat{\boldsymbol{v}}^{(t)}}(\boldsymbol{X})\big)\sum_{j=1}^{D}\sigma'\Big(\sum_{k=1}^{D}\widehat{S}_{j,k}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)},\boldsymbol{X}_k\rangle\Big)\sum_{r=1}^{D}\widehat{S}_{j,r}^{(t)}\langle\boldsymbol{X}_r,\boldsymbol{w}^*\rangle\Big]\Big| \tag{135}$$

$$+D\Big|\mathbb{E}\Big[y\mathfrak{S}\big(-yF_{\widehat{\boldsymbol{A}}^{(t)},\widehat{\boldsymbol{v}}^{(t)}}(\boldsymbol{X})\big)\sum_{j=1}^{D}\sigma'\Big(\sum_{k=1}^{D}\widehat{S}_{j,k}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)},\boldsymbol{X}_k\rangle\Big)\sum_{r=1}^{D}\widehat{S}_{j,r}^{(t)}\langle\boldsymbol{X}_r,\boldsymbol{w}^*\rangle\Big] \tag{136}$$

$$-\mathbb{E}\Big[y\mathfrak{S}\big(-yF_{\widehat{\boldsymbol{A}}^{(t)},\check{\boldsymbol{v}}^{(t)}}(\boldsymbol{X})\big)\sum_{j=1}^{D}\sigma'\Big(\sum_{k=1}^{D}\widehat{S}_{j,k}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_k\rangle\Big)\sum_{r=1}^{D}\widehat{S}_{j,r}^{(t)}\langle\boldsymbol{X}_r,\boldsymbol{w}^*\rangle\Big]\Big| \tag{137}$$

$$+D\Big|\mathbb{E}\Big[y\mathfrak{S}\big(-yF_{\widehat{\boldsymbol{A}}^{(t)},\check{\boldsymbol{v}}^{(t)}}(\boldsymbol{X})\big)\sum_{j=1}^{D}\sigma'\Big(\sum_{k=1}^{D}\widehat{S}_{j,k}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_k\rangle\Big)\sum_{r=1}^{D}\widehat{S}_{j,r}^{(t)}\langle\boldsymbol{X}_r,\boldsymbol{w}^*\rangle\Big] \tag{138}$$

$$-\mathbb{E}\Big[y\mathfrak{S}\big(-yF_{\check{\boldsymbol{A}}^{(t)},\check{\boldsymbol{v}}^{(t)}}(\boldsymbol{X})\big)\sum_{j=1}^{D}\sigma'\Big(\sum_{k=1}^{D}\check{S}_{j,k}^{(t)}\langle\check{\boldsymbol{v}}^{(t)},\boldsymbol{X}_k\rangle\Big)\sum_{r=1}^{D}\check{S}_{j,r}^{(t)}\langle\boldsymbol{X}_r,\boldsymbol{w}^*\rangle\Big]\Big| \tag{139}$$

$$+\varepsilon_{\boldsymbol{v}}^{(t)}\Big|\frac{D}{N}\sum_{i=1}^{N}y[i]\mathfrak{S}\big(-y[i]F_{\widehat{\boldsymbol{A}}^{(t)},\widehat{\boldsymbol{v}}^{(t)}}(\boldsymbol{X}[i])\big)\sum_{j=1}^{D}\sigma'\Big(\sum_{k=1}^{D}\widehat{S}_{j,k}^{(t)}\langle\widehat{\boldsymbol{v}}^{(t)},\boldsymbol{X}_k[i]\rangle\Big)\sum_{r=1}^{D}\widehat{S}_{j,r}^{(t)}\langle\boldsymbol{X}_r[i],\boldsymbol{u}^{(t)}\rangle\Big|. \tag{140}$$

We bound each of the terms above using concentration or lipschitz inequalities. Using the same arguments as in the proof of Lemma E.10, we have $|(133)-(134)|\leqslant \mathrm{poly}(D)/\sqrt{N}$ and $|(136)-(137)|\leqslant \mathrm{poly}(D)\varepsilon_{\boldsymbol{v}}^{(t)}$. Using the same steps as in the proof of Lemma E.14, we have $|(138)-(139)|\leqslant \mathrm{poly}(D)\Delta_{\boldsymbol{A}}^{(t)}$. Lastly, $|(140)|\leqslant \mathrm{poly}(D)\varepsilon_{\boldsymbol{v}}^{(t)}$. Summing up all these terms yields the aimed result.

□

Lemma E.19 shows that we can use the gradient from the semi-idealized process to analyze the dynamics of $\widehat{\alpha}^{(t)}$ in the real process. Therefore, we can derive similar updates for $\widehat{\alpha}^{(t)}$ as in Lemma D.2, Lemma D.8 and Lemma D.11.

Lemma E.20. *Let $\check{\mathcal{T}}_0=\Theta\left(\frac{1}{\eta C(\widehat{\alpha}^{(\mathscr{T})})^{p-1}}\right)$. Therefore, $\widehat{\alpha}^{(t)}$ is updated as*

$$\widehat{\alpha}^{(t+1)}=\widehat{\alpha}^{(t)}+\Theta(\eta C)(\check{G}^{(t)})^p(\widehat{\alpha}^{(t)})^{p-1}.$$

Consequently, $\widehat{\alpha}^{(t)}$ is non-decreasing and after $\check{\mathcal{T}}_0$ iterations, we have $\widehat{\alpha}^{(t)}\geqslant\frac{\Omega(1)}{C^2\lambda_0}$ for $t\geqslant\check{\mathcal{T}}_0$.

Lemma E.21. *Let $\check{\mathcal{T}}_1=\check{\mathcal{T}}_0+\Theta\left(\frac{\lambda_0^p}{\eta e^{\mathrm{polyloglog}(d)}}\right)$. For all $t\in[\check{\mathcal{T}}_1,T]$, we have $\check{\Gamma}^{(t)}\geqslant\frac{\Omega(\lambda_0)}{D}$.*

Lemma E.22. *Let $\check{\mathcal{T}}_1=\check{\mathcal{T}}_0+\Theta\left(\frac{\lambda_0^p}{\eta e^{\mathrm{polyloglog}(d)}}\right)$ and $t\in[\check{\mathcal{T}}_1,T]$. $\widehat{\alpha}^{(t)}$ updates as*

$$\widehat{\alpha}^{(t+1)}=\widehat{\alpha}^{(t)}+\Theta(\eta C)(\check{G}^{(t)})^p(\widehat{\alpha}^{(t)})^{p-1}. \tag{141}$$

Consequently, $\widehat{\alpha}^{(t)}$ is non-decreasing and eventually $\widehat{\alpha}^{(T)}=\frac{\mathrm{polylog}(d)}{C^2\lambda_0}$.

Auxiliary lemma. The following lemma is useful to prove Lemma E.10 and Lemma E.14.

Lemma E.23. *The time at which $\widehat{\alpha}^{(t)}$ stops increasing is $\check{\mathcal{T}}_2=\check{\mathcal{T}}_1+\Theta\left(\frac{\lambda_0^{p-2}}{\eta\mathrm{polylog}(d)}\right)$.*

Proof of Lemma E.23. The result is obtained by applying Lemma K.1 to (141). We have:

$$\check{\mathcal{T}}_2=\check{\mathcal{T}}_1+\frac{3(\lambda_0)^{p-2}}{\eta\mathrm{polylog}(d)e^{\mathrm{polyloglog}(d)}}+\frac{2^p\lambda_0^p}{e^{\mathrm{polyloglog}(d)}}\log\log(d).$$

□

E.2.7 The realistic model fits the labeling function

Lemma E.24. *In the realistic case, the model fits the labeling function i.e.*

$$\mathbb{P}_{\mathcal{D}}[f^*(\boldsymbol{X})F_{\widehat{\boldsymbol{A}}^{(T)},\widehat{\boldsymbol{v}}^{(T)}}(\boldsymbol{X}) > 0] \geqslant 1 - o(1). \tag{142}$$

Proof of Lemma E.24. We bound the population risk $\mathcal{L}(\widehat{\boldsymbol{A}}^{(T)}, \widehat{\boldsymbol{v}}^{(T)})$. We have:

$$\mathcal{L}(\widehat{\boldsymbol{A}}^{(T)}, \widehat{\boldsymbol{v}}^{(T)}) \leqslant \left|\mathcal{L}(\widehat{\boldsymbol{A}}^{(T)}, \widehat{\boldsymbol{v}}^{(T)}) - \mathcal{L}(\check{\boldsymbol{A}}^{(T)}, \check{\boldsymbol{v}}^{(T)})\right| + \mathcal{L}(\check{\boldsymbol{A}}^{(T)}, \check{\boldsymbol{v}}^{(T)}). \tag{143}$$

Using Lemma E.25, we have $\mathcal{L}(\check{\boldsymbol{A}}^{(T)}, \check{\boldsymbol{v}}^{(T)}) \leqslant 1/\mathrm{poly}(d)$. We now bound the first summand in (143) using the 1-Lipschitzness of the logistic function and get:

$$\left|\mathcal{L}(\widehat{\boldsymbol{A}}^{(T)}, \widehat{\boldsymbol{v}}^{(T)}) - \mathcal{L}(\check{\boldsymbol{A}}^{(T)}, \check{\boldsymbol{v}}^{(T)})\right| \leqslant \mathbb{E}\Big[\left|F_{\widehat{\boldsymbol{A}}^{(T)},\widehat{\boldsymbol{v}}^{(T)}}(\boldsymbol{X}) - F_{\check{\boldsymbol{A}}^{(T)},\check{\boldsymbol{v}}^{(T)}}(\boldsymbol{X})\right|\Big] \tag{144}$$

Using Lemma E.26, we have $\mathbb{E}\Big[\left|F_{\widehat{\boldsymbol{A}}^{(T)},\widehat{\boldsymbol{v}}^{(T)}}(\boldsymbol{X}) - F_{\check{\boldsymbol{A}}^{(T)},\check{\boldsymbol{v}}^{(T)}}(\boldsymbol{X})\right|\Big] \leqslant 1/\mathrm{poly}(d)$. Therefore, we deduce that $\mathcal{L}(\widehat{\boldsymbol{A}}^{(T)}, \widehat{\boldsymbol{v}}^{(T)}) \leqslant 1/\mathrm{poly}(d)$. Since the 0-1 loss is a convex surrogate, we have

$$\mathbb{P}_{(\boldsymbol{X},y)\sim\mathcal{D}}[yF_{\widehat{\boldsymbol{A}}^{(T)},\widehat{\boldsymbol{v}}^{(T)}}(\boldsymbol{X}) < 0] \leqslant \mathcal{L}(\widehat{\boldsymbol{A}}^{(T)}, \widehat{\boldsymbol{v}}^{(T)}) \leqslant 1/\mathrm{poly}(d). \tag{145}$$

We can further expand (145) as in the proof of Lemma D.15 and deduce the aimed result. □

To prove Lemma E.24, we use the following auxiliary lemma.

Lemma E.25. *After T iterations, the population risk in the semi-idealized case converges i.e. $\check{\mathcal{L}}(\check{\boldsymbol{A}}^{(T)}, \check{\boldsymbol{v}}^{(T)}) \leqslant o(1)$.*

Proof of Lemma E.25. The proof is similar to the one of Lemma D.15. □

Lemma E.26. *For all $\boldsymbol{X}$ sampled from $\mathcal{D}$, we have*

$$\left|F_{\widehat{\boldsymbol{A}}^{(T)},\widehat{\boldsymbol{v}}^{(T)}}(\boldsymbol{X}) - F_{\check{\boldsymbol{A}}^{(T)},\check{\boldsymbol{v}}^{(T)}}(\boldsymbol{X})\right| \leqslant 1/\mathrm{poly}(d).$$

Proof of Lemma E.26. We have:

$$\left|F_{\widehat{\boldsymbol{A}}^{(T)},\widehat{\boldsymbol{v}}^{(T)}}(\boldsymbol{X}) - F_{\boldsymbol{Q}^{(T)},\check{\boldsymbol{v}}^{(T)}}(\boldsymbol{X})\right| \leqslant \left|F_{\widehat{\boldsymbol{A}}^{(T)},\widehat{\boldsymbol{v}}^{(T)}}(\boldsymbol{X}) - F_{\widehat{\boldsymbol{A}}^{(T)},\check{\boldsymbol{v}}^{(T)}}(\boldsymbol{X})\right| \tag{146}$$

$$+ \left|F_{\widehat{\boldsymbol{A}}^{(T)},\check{\boldsymbol{v}}^{(T)}}(\boldsymbol{X}) - F_{\check{\boldsymbol{A}}^{(T)},\check{\boldsymbol{v}}^{(T)}}(\boldsymbol{X})\right|. \tag{147}$$

We now separately bound (146) and (147).

Bound on (146). Since $x \mapsto x^p + \nu x$ is Lipschitz on a bounded domain, we have:

$$(146) \leqslant \sum_{m=1}^{D} \sigma'\Big(D\sum_{r=1}^{D} \widehat{S}_{m,r}^{(T)}|\langle\widehat{\boldsymbol{v}}^{(T)}, \boldsymbol{X}_r\rangle|\Big) D\sum_{r=1}^{D} \widehat{S}_{m,r}^{(T)}|\langle\widehat{\boldsymbol{v}}^{(T)} - \check{\boldsymbol{v}}^{(T)}, \boldsymbol{X}_r\rangle| \tag{148}$$

$$\leqslant \varepsilon_{\boldsymbol{v}}^{(T)} \sum_{m=1}^{D} \sigma'\Big(D\sum_{r=1}^{D} \widehat{S}_{m,r}^{(T)}|\langle\widehat{\boldsymbol{v}}^{(T)}, \boldsymbol{X}_r\rangle|\Big) D\sum_{r=1}^{D} \widehat{S}_{m,r}^{(T)}|\langle\boldsymbol{u}^{(t)}, \boldsymbol{X}_r\rangle|, \tag{149}$$

since $\langle\widehat{\boldsymbol{v}}^{(T)}, \boldsymbol{w}^*\rangle = \langle\check{\boldsymbol{v}}^{(T)}, \boldsymbol{w}^*\rangle$. Using Cauchy-Schwarz inequality, (148) simplifies as:

$$(146) \leqslant \sum_{m=1}^{D} \sigma'\Big(D\sum_{r=1}^{D} \widehat{S}_{m,r}^{(T)}|\langle\widehat{\boldsymbol{v}}^{(T)}, \boldsymbol{X}_r\rangle|\Big)^{p-1} D\sum_{r=1}^{D} \widehat{S}_{m,r}^{(T)}\varepsilon_{\boldsymbol{v}}^{(T)} \leqslant \mathrm{poly}(D)\varepsilon_{\boldsymbol{v}}^{(T)}. \tag{150}$$

Using Lemma E.10, we conclude that (146) $\leqslant 1/\mathrm{poly}(d)$.

Bound on (147). We again use the Lipschitzness of the power function and get:

$$(147) \leqslant \sum_{m=1}^{D} \sigma'\Big(D\sum_{r=1}^{D} \max\{\check{S}_{m,r}^{(T)}, \widehat{S}_{m,r}^{(T)}\}|\langle\boldsymbol{v}^{(T)}, \boldsymbol{X}_r\rangle|\Big) D\sum_{r=1}^{D} |\widehat{S}_{m,r}^{(T)} - \check{S}_{m,r}^{(T)}| \cdot |\langle\widehat{\boldsymbol{v}}^{(T)}, \boldsymbol{X}_r\rangle|. \tag{151}$$

We apply Lemma E.17 in (151) to get (147) $\leqslant \mathrm{poly}(D)\Delta_{\boldsymbol{A}}^{(T)}$. Finally, we apply Lemma E.14 to get (147) $\leqslant 1/\mathrm{poly}(d)$.

□

F Transfer Learning

In this section, we show that a transformer that has been pre-trained on a structured dataset require a few samples to generalize in a new dataset sharing the same structure.

Theorem 5.1. *Let $\widehat{\boldsymbol{A}}$ be the attention matrix obtained after pre-training as in subsection 4.2. Assume that we run GD for T iterations on ($\widetilde{\text{E}}$) to fine-tune the value vector. Using $\widetilde{N} \leqslant \text{polylog}(D)$ samples, the model (T) transfers to $\widetilde{\mathcal{D}}$ i.e. $\mathbb{P}_{\widetilde{\mathcal{D}}}[f^*(\boldsymbol{X})F_{\widehat{\boldsymbol{A}},\widetilde{\boldsymbol{v}}^{(T)}}(\boldsymbol{X}) > 0] \geqslant 1 - o(1)$.*

Proof of Theorem 5.1. Actually, even one step of the update using normalized gradient descent on $\boldsymbol{v}$ can already achieve test accuracy $\geqslant 1 - o(1)$. We know that for a datum $(\boldsymbol{X}, y)$, the gradient of $L(\boldsymbol{X})$ with respect to $\boldsymbol{v}$ is

$$\nabla_{\boldsymbol{v}} L(\boldsymbol{X}) = -y\mathfrak{S}(-yF(\boldsymbol{X})) \sum_{m=1}^{D} \sigma'(\langle \boldsymbol{O}_m^{(t)}, \widetilde{\boldsymbol{v}}^{(t)} \rangle) \boldsymbol{O}_m^{(t)} \tag{152}$$

Since $\widetilde{\boldsymbol{v}}^{(0)} = \boldsymbol{0}$, we have $F_{\widetilde{\boldsymbol{v}}^{(0)}}(\boldsymbol{X}) = 0$ and $\sigma'(\langle \boldsymbol{O}_m^{(0)}, \widetilde{\boldsymbol{v}}^{(0)} \rangle) = \nu$. Thus, the gradient (152) simplifies to

$$\nabla_{\boldsymbol{v}} L(\boldsymbol{X}) = -\frac{\nu y}{2} \sum_{j=1}^{D} \boldsymbol{X}_j \sum_{m=1}^{D} \widehat{S}_{m,j}^{(t)} \tag{153}$$

By symmetry of the $\widehat{S}_{j,m}^{(t)}$, we know that

$$\Big| \sum_{m=1}^{D} \widehat{S}_{m,j}^{(t)} - D \Big| \leqslant \frac{D}{\text{poly}(d)}, \tag{154}$$

where $\frac{1}{\text{poly}(d)}$ comes from the $\varepsilon_{\boldsymbol{Q}}$ part in the previous section. Moreover, since the noise and feature noise has mean zero independent of y, we know that there exists some value $c_0 > 0$ (roughly equal to αDC)) such that:

$$\mathbb{E}[\nabla_{\boldsymbol{v}} L(\boldsymbol{X})] = c_0 \widetilde{\boldsymbol{w}}^*. \tag{155}$$

Now, by standard concentration inequality, we know that for N i.i.d. samples $\boldsymbol{X}[i], y[i]$, with high probability

$$\boldsymbol{\nabla}_0 := \sum_{i=1}^{N} \nabla_{\boldsymbol{v}} L(\boldsymbol{X}[i]) = c_0 \widetilde{\boldsymbol{w}}^* + \varepsilon_0 \widetilde{\boldsymbol{w}}^* + \boldsymbol{\chi}_0, \tag{156}$$

where ε_0 comes from the feature noise

$$|\varepsilon_0| \leqslant \frac{c_0 q D \log(d)}{\sqrt{N}}, \tag{157}$$

and $\boldsymbol{\chi}_0$ comes from the noise:

$$\|\boldsymbol{\chi}_0\|_2 \leqslant \frac{c_0 \sigma \sqrt{d} \sqrt{D}}{\sqrt{N}}. \tag{158}$$

Therefore, if we update using normalized GD:

$$\widetilde{\boldsymbol{v}}^{(1)} = \widetilde{\boldsymbol{v}}^{(0)} + \frac{\boldsymbol{\nabla}_0}{\|\boldsymbol{\nabla}_0\|_2} \tag{159}$$

we have that:

$$\widetilde{\boldsymbol{v}}^{(1)} = c_1 \widetilde{\boldsymbol{w}}^* + \boldsymbol{\chi}_1, \tag{160}$$

where $\boldsymbol{\chi}_1 \perp \widetilde{\boldsymbol{w}}^*$, $\|\boldsymbol{\chi}_1\|_2 = O(1)$ and $c_1 \geqslant \tilde{\Omega}(1)/\sqrt{D}$. Now, for a new datum $\boldsymbol{X}^{new}$ with noises $\{\boldsymbol{\xi}_i^{new}\}_{i=1}^{D}$, we know that w.h.p

$$|\langle \boldsymbol{\xi}_i^{new}, \widetilde{\boldsymbol{v}}^{(1)} \rangle| \leqslant \frac{\tilde{O}(1)}{\sqrt{d}}, \qquad |\langle \widetilde{\boldsymbol{w}}^*, \widetilde{\boldsymbol{v}}^{(1)} \rangle| \geqslant \frac{\tilde{\Omega}(1)}{\sqrt{D}}. \tag{161}$$

We can prove the test accuracy is small using the same proof as in Lemma D.15, where we show that:

$$y\sigma(\langle \widetilde{\boldsymbol{v}}^{(1)}, \boldsymbol{O}_m^{(1)} \rangle) \geqslant 0, \tag{162}$$

for $m \in \mathcal{S}_{\ell(\boldsymbol{X})}$ and it dominates the other $y\sigma(\langle \widetilde{\boldsymbol{v}}^{(1)}, \boldsymbol{O}_j^{(1)} \rangle)$ for $j \in \mathcal{S}_\ell$ with $\ell \neq \ell(\boldsymbol{X})$. Therefore, we prove that $\mathbb{P}_{\widetilde{\mathcal{D}}}[yF(\boldsymbol{X}) > 0] \geqslant 1 - o(1)$ which implies the aimed result. □

Theorem 5.2. *Let $\mathcal{A}\colon \mathbb{R}^{d\times D} \to \{-1, 1\}$ be a binary classification algorithm without patch association knowledge. Then, it needs $D^{\Omega(1)}$ training samples to get test error $\leqslant o(1)$ on $\widetilde{\mathcal{D}}$.*

Proof of Theorem 5.2. Let $\mathcal{A}$ be an algorithm. Assume that at training time the algorithm $\mathcal{A}$ has access to $D^{o(1)}$ training data, Since each input X is made of D patches, this means that there exist at least $\Omega(1)$ fraction of $k \in [D]$ such that $\mathcal{A}$ has not seen training samples with $\ell(X) = k$. Consider the following two distributions over $\{-1, 0, 1\}^m$:

1. $\mathcal{D}_1$: Sample $z \in \{0, 1\}^m$ where each z_i i.i.d. 1 w.p. $q/2$, -1 w.p. $q/2$ and 0 otherwise.
2. $\mathcal{D}_2$: Sample a set $\mathcal{S}$ uniformly at random from $[m]$ of size C, set all $z_i = 1$ for $i \in \mathcal{S}$, and sample other z_j i.i.d. 1 w.p. $q/2$, -1 w.p. $q/2$ and 0 otherwise.

We can easily see that as long as $qm = \text{poly}(C)$, then

$$\mathbf{TV}(\mathcal{D}_1, \mathcal{D}_2) = o(1)$$

This implies that $\mathcal{A}$ must have bad generalization error ($\Omega(1)$) on $\widetilde{\mathcal{D}}$.

□

G Gradient descent updates in the idealized process

In this section, we derive the gradient descent updates of $A_{i,j}$ in the idealized learning process.

G.1 Indices in the same set: $i, j \in \mathcal{S}_\ell$

Lemma G.1. *Let $T > 0$ be the time where the population loss is at most $o(1)$ and $t \in [0, T]$. Then, $\gamma^{(t)}$ satisfies the update*

$$\gamma^{(t+1)} = \gamma^{(t)} + \eta\Theta(C)\alpha^{(t)}\Big(D\alpha^{(t)}\big[\Lambda^{(t)} + (C-1)\Gamma^{(t)}\big]\Big)^{p-1}\Gamma^{(t)}.$$

Proof of Lemma G.1. Let $\ell \in [L]$ and $i, j \in \mathcal{S}_\ell$ with $i \neq j$. The main idea of the proof is to bound the gradient of $L(\boldsymbol{X})$ with respect to $A_{i,j}$. This gradient is given by Lemma H.2 and is made of two terms: the σ' term and the sum outside σ'. We distinguish the following cases and bound these two terms.

1. $\ell = \ell(\boldsymbol{X})$. We first bound the outside sum. Using Lemma K.3, we have

$$\Big|\sum_{h\neq\ell(\boldsymbol{X})}\sum_{r\in\mathcal{S}_h}(1 - y\delta_r) - (D - C)\Big| \leqslant \sum_{h\neq\ell(\boldsymbol{X})}\sum_{r\in\mathcal{S}_h}|\delta_r| \leqslant qD\log(d). \tag{163}$$

Since $(D - C)/2 - qD\log(d) \geqslant 0$, we rewrite (163) as:

$$\frac{1}{2}(D - C) \leqslant \sum_{h\neq\ell(\boldsymbol{X})}\sum_{r\in\mathcal{S}_h}(1 - y\delta_r) \leqslant 2(D - C). \tag{164}$$

Regarding the sum inside σ', we use Lemma G.3 which shows:

$$\Lambda^{(t)} + (C-1)\Gamma^{(t)} + \Xi^{(t)}\sum_{h\neq\ell(\boldsymbol{X})}\sum_{r\in\mathcal{S}_h}\delta_{h,r} = \Theta(1)\left(\Lambda^{(t)} + (C-1)\Gamma^{(t)}\right). \tag{165}$$

By using (164) and (165), we finally obtain:

$$-\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}} = \Theta(D\alpha^{(t)})\mathfrak{S}(-yF(\boldsymbol{X}))\left(D\alpha^{(t)}(\Lambda^{(t)} + (C-1)\Gamma^{(t)})\right)^{p-1}\Gamma^{(t)}(D-C)\Xi^{(t)}. \quad (166)$$

2. $\ell \neq \ell(\boldsymbol{X})$ and $\delta_s = 0$ for all $s \in \mathcal{S}_\ell$. We first bound the outside sum. Since $\delta_s = 0$, the only non-zero term is the one with factor $\Xi^{(t)}$. Using triangle inequality, we have:

$$\Xi^{(t)}\Big|Cy + \sum_{h\neq\{\ell(\boldsymbol{X}),\ell\}}\sum_{r\in\mathcal{S}_h}\delta_r\Big| \leqslant \Xi^{(t)}\left(C + \sum_{h\neq\{\ell(\boldsymbol{X}),\ell\}}\sum_{r\in\mathcal{S}_h}|\delta_r|\right) = \Xi^{(t)}\kappa(\boldsymbol{X}). \quad (167)$$

We now bound the sum inside σ'. This sum is actually equal to the outside sum and we can therefore use the bound (167). Therefore, the overall gradient is bounded as:

$$\Big|\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\Big| \leqslant \Theta(D\alpha^{(t)})\mathfrak{S}(-yF(\boldsymbol{X}))\left(D\Xi^{(t)}\alpha^{(t)}\kappa(\boldsymbol{X})\right)^{p-1}\Gamma^{(t)}\Xi^{(t)}\kappa(\boldsymbol{X}). \quad (168)$$

3. $\ell \neq \ell(\boldsymbol{X})$ and at least one $\delta_s \neq 0$ and $\delta_i = 0$. We first bound the outside sum. Using Lemma K.3 and $\Lambda^{(t)} + \Gamma^{(t)} + \Xi^{(t)} = 1$, we have:

$$\begin{aligned}&\alpha^{(t)}\Big|\Lambda^{(t)}\delta_j + \Gamma^{(t)}\sum_{r\in\mathcal{S}_\ell\setminus\{i\}}(\delta_j-\delta_r) + \Xi^{(t)}C(\delta_j - y) + \Xi^{(t)}\sum_{h\neq\{\ell(\boldsymbol{X}),\ell\}}\sum_{m\in\mathcal{S}_h}(\delta_j-\delta_m)\Big| \\ \leqslant&\alpha^{(t)}\left[(C-1)\Gamma^{(t)} + \Xi^{(t)}\left(C + \Theta(qD)\log(d)\right) + \mathbf{1}_{\delta_j\neq 0}\right].\end{aligned} \quad (169)$$

We lastly apply Induction Hypothesis B.1 to show that (169) is less or equal to $\Theta(\alpha^{(t)})$. We now bound the sum inside σ'.

$$\begin{aligned}\alpha^{(t)}\Big|\Gamma^{(t)}\sum_{r\in\mathcal{S}_\ell\setminus\{i,j\}}\delta_r + \Xi^{(t)}\Big(Cy + \sum_{h\neq\{\ell(\boldsymbol{X}),\ell\}}\sum_{m\in\mathcal{S}_h}\delta_m\Big)\Big| &\leqslant \alpha^{(t)}\Big(\Gamma^{(t)}\sum_{r\in\mathcal{S}_\ell\setminus\{i,j\}}|\delta_r| + \Xi^{(t)}\kappa(\boldsymbol{X})\Big) \\ &\leqslant \alpha^{(t)}\left((C-1)\Gamma^{(t)} + \Xi^{(t)}\kappa(\boldsymbol{X})\right).\end{aligned} \quad (170)$$

The overall bound on the derivative is:

$$\Big|\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\Big| \leqslant \Theta(D\alpha^{(t)})\mathfrak{S}(-yF(\boldsymbol{X}))\Big(D\alpha^{(t)}\big((C-1)\Gamma^{(t)} + \Xi^{(t)}\kappa(X)\big)\Big)^{p-1}\Gamma^{(t)}. \quad (171)$$

4. $\ell \neq \ell(\boldsymbol{X})$ where $\delta_s \neq 0$ for all s. We first bound the outside sum as follows.

$$\begin{aligned}&\alpha^{(t)}\Big|\Lambda^{(t)}(\delta_j-\delta_i) + \Gamma^{(t)}\sum_{r\in\mathcal{S}_\ell\setminus\{i,j\}}(\delta_j-\delta_r) + \Xi^{(t)}\Big[C(\delta_j - y) + \sum_{h\neq\{\ell(\boldsymbol{X}),\ell\}}\sum_{m\in\mathcal{S}_h}(\delta_j-\delta_m)\Big]\Big| \\ &\leqslant \alpha^{(t)}(\Lambda^{(t)} + (C-1)\Gamma^{(t)} + \Xi^{(t)}\kappa(\boldsymbol{X}) + \mathbf{1}_{\delta_j\neq 0}).\end{aligned} \quad (172)$$

We now bound the sum inside σ'.

$$\begin{aligned}&\alpha^{(t)}\Big|\Lambda^{(t)}\delta_i + \Gamma^{(t)}\sum_{r\in\mathcal{S}_\ell\setminus\{i,j\}}\delta_r + \Xi^{(t)}\Big[Cy + \sum_{h\neq\{\ell(\boldsymbol{X}),\ell\}}\sum_{m\in\mathcal{S}_h}\delta_m\Big]\Big| \\ \leqslant&\alpha^{(t)}\Big(\Lambda^{(t)} + (C-1)\Gamma^{(t)} + \Xi^{(t)}\kappa(\boldsymbol{X})\Big).\end{aligned} \quad (173)$$

We lastly apply Lemma G.3 to show that (173) is bounded by $\Theta(\alpha^{(t)})(\Lambda^{(t)} + (C-1)\Gamma^{(t)})$. Therefore, the overall gradient is bounded as

$$\begin{aligned}\Big|\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\Big| \leqslant \Theta(D\alpha^{(t)})\mathfrak{S}(-yF(\boldsymbol{X}))&\left(D\alpha^{(t)}((C-1)\Gamma^{(t)} + \Lambda^{(t)})\right)^{p-1}\Gamma^{(t)}\cdot \\ &\left(\Lambda^{(t)} + (C-1)\Gamma^{(t)} + \Xi^{(t)}\kappa(\boldsymbol{X}) + \mathbf{1}_{\delta_j\neq 0}\right).\end{aligned} \quad (174)$$

Putting all the pieces together. We now bound the derivative of the population loss. Using Tower property and Lemma I.1, we have:

$$\mathbb{E}_{\boldsymbol{X}}\left[\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\right] = \mathbb{E}_{\boldsymbol{X}}\left[\frac{\partial L(\pi(\boldsymbol{X}))}{\partial A_{i,j}}\right] = \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{\pi_1,\pi_2}\Big[\frac{\partial L(\pi(\boldsymbol{X}))}{\partial A_{i,j}}\Big|\boldsymbol{X}\right]\Big]. \tag{175}$$

For a *fixed* $\boldsymbol{X}$, we now bound the derivative of the loss evaluated in $\boldsymbol{X}$. For $i, j \in \mathcal{S}_\ell$, we distinguish the four possible events depending on the randomness of π.

- **Event a**: "$\ell = \ell(\pi(\boldsymbol{X}))$" occurs with probability $1/L$.
- **Event b**: "$\ell \neq \ell(\pi(\boldsymbol{X}))$ and $\delta_s = 0$ for all $s \in \mathcal{S}_\ell$" occurs with probability $(1-1/L)(1-q)^C$.
- **Event $\mathbf{c_k}$**: "$\ell \neq \ell(\pi(\boldsymbol{X}))$ and $\#\{s : \delta_s \neq 0\} = k$ for $1 \leqslant k \leqslant C-1$ and $\delta_i = 0$" occurs with probability $(1-1/L)\binom{C-1}{k}q^k(1-q)^{C-k}$.
- **Event d**: "$\ell \neq \ell(\pi(\boldsymbol{X}))$ and $\delta_s \neq 0$ for all s" occurs with probability $(1-1/L)\cdot q^C$.

Therefore, the derivative of the loss in $\boldsymbol{X}$ is:

$$\begin{aligned}\mathbb{E}_{\pi_1,\pi_2}\Big[\frac{\partial L(\pi(\boldsymbol{X}))}{\partial A_{i,j}}\Big|\boldsymbol{X}\Big] &= \frac{1}{L}\mathbb{E}_{\pi_1,\pi_2}\Big[\frac{\partial L(\pi(\boldsymbol{X}))}{\partial A_{i,j}}\;\Big|\;\boldsymbol{X},\mathbf{a}\Big] \\ &+ \left(1-\frac{1}{L}\right)(1-q)^C\mathbb{E}_{\pi_1,\pi_2}\Big[\frac{\partial L(\pi(\boldsymbol{X}))}{\partial A_{i,j}}\;\Big|\;\boldsymbol{X},\mathbf{b}\Big] \\ &+ \left(1-\frac{1}{L}\right)\sum_{k=1}^{C-1}\binom{C-1}{k}q^k(1-q)^{C-k}\mathbb{E}_{\pi_1,\pi_2}\Big[\frac{\partial L(\pi(\boldsymbol{X}))}{\partial A_{i,j}}\;\Big|\;\boldsymbol{X},\mathbf{c}_k\Big] \\ &+ \left(1-\frac{1}{L}\right)q^C\mathbb{E}_{\pi_1,\pi_2}\Big[\frac{\partial L(\pi(\boldsymbol{X}))}{\partial A_{i,j}}\;\Big|\;\boldsymbol{X},\mathbf{d}\Big].\end{aligned} \tag{176}$$

Event a is the event that is the most likely to happen. Therefore, we only take into account $\mathbb{E}_{\pi_1,\pi_2}\left[\frac{\partial L(\pi(\boldsymbol{X}))}{\partial A_{i,j}}\;\Big|\;\mathbf{a}\right]$ in (176) and obtain:

$$\mathbb{E}_{\boldsymbol{X}}\left[\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\right] = \frac{\Theta(D\alpha^{(t)})}{L}\mathbb{E}_X[\mathfrak{S}(-yF(\boldsymbol{X}))]\left(D\alpha^{(t)}(\Lambda^{(t)}+(C-1)\Gamma^{(t)}))\right)^{p-1}\Gamma^{(t)}. \tag{177}$$

Since the population loss is a $\Omega(1)$ for $t \leqslant T$, this implies that $\mathbb{E}_{\boldsymbol{X}}[\mathfrak{S}(-yF(\boldsymbol{X}))] = \Theta(1)$ (Lemma D.10). We thus plug (177) in the update of $\gamma^{(t)}$ to obtain the desired result. □

Corollary G.1. *Let $T > 0$ be the time where the population loss is $o(1)$ and $t \leqslant T$. Let $G^{(t)} := D(\Lambda^{(t)} + (C-1)\Gamma^{(t)})$. The update of $\gamma^{(t)}$ satisfies:*

$$\gamma^{(t+1)} = \gamma^{(t)} + \Theta(C\eta)(\alpha^{(t)})^p\Gamma^{(t)}(G^{(t)})^{p-1}.$$

Proof of Corollary G.1. Lemma G.1 provides the update rule of $\gamma^{(t)}$.

$$\gamma^{(t+1)} = \gamma^{(t)} + \eta\Theta(C)(\alpha^{(t)})^p\Gamma^{(t)}(G^{(t)})^{p-1}.$$

□

G.2 Update for $i \in \mathcal{S}_\ell$ and $j \in \mathcal{S}_m$

Lemma G.2. *Let $T > 0$ be the time where the population loss is at most $o(1)$ and $t \in [0, T]$. Then, $\rho^{(t)}$ satisfies the update*

$$\begin{aligned}\left|\frac{\rho^{(t+1)}-\rho^{(t)}}{\eta}\right| &\leqslant \Theta(C^2\alpha^{(t)})\left(D\alpha^{(t)}(\Lambda^{(t)}+(C-1)\Gamma^{(t)})\right)^{p-1}\Xi^{(t)}\frac{\lambda_0}{D} \\ &+ \Theta(\alpha^{(t)})\left(D\alpha^{(t)}qD\log(d)\Xi^{(t)}\right)^{p-1}\Xi^{(t)}qD\log(d).\end{aligned}$$

Proof of Lemma G.1. Let $\ell, m \in [L]$ such that $\ell \neq m$ and $i \in \mathcal{S}_\ell$, $j \in \mathcal{S}_m$. The main idea of the proof is to bound the gradient of $L(\boldsymbol{X})$ with respect to $A_{i,j}$. This gradient is given by Lemma H.2 and is made of two terms: the σ' term and the sum outside σ'. We distinguish the following cases and bound these two terms.

1. $\ell = \ell(\boldsymbol{X})$ and $\delta_j = 0$. We first bound the outside sum. We apply Lemma G.3 to obtain:

$$\alpha^{(t)}\left|\left(\Lambda^{(t)} + (C-1)\Gamma^{(t)}\right)y + \Xi^{(t)} \sum_{h \neq \ell(\boldsymbol{X})} \sum_{r \in \mathcal{S}_h} \delta_r\right| \leqslant 2\alpha^{(t)}\left(\Lambda^{(t)} + (C-1)\Gamma^{(t)}\right). \tag{178}$$

We now bound the sum inside σ'. This sum is actually equal to the outside sum and we can therefore use the bound (178). Therefore, the overall gradient is bounded as:

$$\left|\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\right| \leqslant \Theta(D\alpha^{(t)})\mathfrak{S}(-yF(\boldsymbol{X}))\left(D\alpha^{(t)}(\Lambda^{(t)} + (C-1)\Gamma^{(t)})\right)^{p-1}\Xi^{(t)}(\Lambda^{(t)} + (C-1)\Gamma^{(t)}). \tag{179}$$

2. $\ell = \ell(\boldsymbol{X})$ and $\delta_j \neq 0$. We first bound the outside sum. We successively apply $\Lambda^{(t)} + \Gamma^{(t)} + \Xi^{(t)} = 1$, Lemma K.3 and Induction Hypothesis B.1 to obtain:

$$\begin{aligned}
&\alpha^{(t)}\left|\Lambda^{(t)}(\delta_j - y) + (C-1)\Gamma^{(t)}(\delta_j - y) + \Xi^{(t)} \sum_{h \neq \ell(\boldsymbol{X})} \sum_{r \in \mathcal{S}_h} (\delta_j - \delta_r)\right| \\
&\leqslant \alpha^{(t)}\left[\Lambda^{(t)} + (C-1)\Gamma^{(t)} + \Xi^{(t)} \sum_{h \neq \ell(\boldsymbol{X})} \sum_{r \in \mathcal{S}_h} |\delta_r| + \mathbf{1}_{\delta_j \neq 0}\right] \\
&\leqslant \alpha^{(t)}\left[\Lambda^{(t)} + (C-1)\Gamma^{(t)} + \Xi^{(t)} qD\log(d) + \mathbf{1}_{\delta_j \neq 0}\right] \\
&\leqslant \alpha^{(t)}\left[\frac{e^{\mathrm{polyloglog}(d)} + \lambda_0}{D} + \Theta(q\log(d)) + \mathbf{1}_{\delta_j \neq 0}\right] = \Theta(\alpha^{(t)}).
\end{aligned} \tag{180}$$

We now bound the sum inside σ'. We successively apply Lemma K.3, triangle inequality and Lemma G.3 to obtain:

$$\alpha^{(t)}\left|y(\Lambda^{(t)} + (C-1)\Gamma^{(t)}) + \Xi^{(t)} \sum_{h \neq \ell(\boldsymbol{X})} \sum_{r \in \mathcal{S}_h} \delta_r\right| \leqslant 2\alpha^{(t)}(\Lambda^{(t)} + (C-1)\Gamma^{(t)}). \tag{181}$$

Thus, we use (180) and (181) to obtain a bound on the derivative.

$$\left|\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\right| \leqslant \Theta(D\alpha^{(t)})\mathfrak{S}(-yF(\boldsymbol{X}))\left(D\alpha^{(t)}(\Lambda^{(t)} + (C-1)\Gamma^{(t)})\right)^{p-1}\Xi^{(t)}. \tag{182}$$

3. $\ell \neq \ell(\boldsymbol{X})$ and $\delta_s = 0$ for all s and $m = \ell(\boldsymbol{X})$. We first bound the outside sum. Since $\Lambda^{(t)} + (C-1)\Gamma^{(t)} + (D-C)\Xi^{(t)} = 1$, we have:

$$\Xi^{(t)}\left|y\left(\Lambda^{(t)} + (C-1)\Gamma^{(t)}\right) + \Xi^{(t)} \sum_{h \neq \{\ell(\boldsymbol{X}), \ell\}} \sum_{r \in \mathcal{S}_h} (y - \delta_r)\right| \leqslant \Xi^{(t)}\left(\Xi^{(t)}\kappa(\boldsymbol{X}) + \mathbf{1}_{m = \ell(\boldsymbol{X})}\right). \tag{183}$$

We now bound the sum inside σ'.

$$\Xi^{(t)}\left|y + \sum_{h \neq \{\ell(\boldsymbol{X}), \ell\}} \sum_{r \in \mathcal{S}_h} \delta_r(X)\right| \leqslant \Xi^{(t)}\kappa(\boldsymbol{X}). \tag{184}$$

Using (183) and (184), we obtain a bound on the derivative.

$$\left|\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\right| \leqslant \Theta(D\alpha^{(t)})\mathfrak{S}(-yF(\boldsymbol{X}))\left(D\alpha^{(t)}\kappa(\boldsymbol{X})\Xi^{(t)}\right)^{p-1}\Xi^{(t)}\left(\Xi^{(t)}\kappa(\boldsymbol{X}) + \mathbf{1}_{m = \ell(\boldsymbol{X})}\right). \tag{185}$$

4. $\ell \neq \ell(\boldsymbol{X})$ and $\delta_s \neq 0$ for some s and $m = \ell(\boldsymbol{X})$. We first bound the outside sum. We apply $\Lambda^{(t)} + (C-1)\Gamma^{(t)} + (D-C)\Xi^{(t)} = 1$ and Lemma G.4 to get:

$$\begin{aligned}
&\Xi^{(t)} \left| \Lambda^{(t)}(y - \delta_i) + \Gamma^{(t)} \sum_{r \in \mathcal{S}_\ell \setminus \{i\}} (y - \delta_r) + \Xi^{(t)} \sum_{h \neq \{\ell(X), \ell\}} \sum_{r \in \mathcal{S}_h} (y - \delta_r) \right| \\
&\leqslant \Xi^{(t)} \left(\Lambda^{(t)} + (C-1)\Gamma^{(t)} + \Xi^{(t)} \sum_{h \neq \{\ell(\boldsymbol{X}), \ell\}} \sum_{r \in \mathcal{S}_h} |\delta_r| + \mathbf{1}_{m = \ell(\boldsymbol{X})} \right) \\
&\leqslant \Theta(\Xi^{(t)}).
\end{aligned} \tag{186}$$

We now bound the sum inside σ'. We apply Lemma G.3 to obtain:

$$\begin{aligned}
&\left| \Lambda^{(t)} \delta_i(X) + \Gamma^{(t)} \sum_{r \in \mathcal{S}_\ell \setminus \{i\}} \delta_r(X) + \Xi^{(t)} \left(Cy + \sum_{h \in [L] \setminus \{\ell(\boldsymbol{X}), \ell\}} \sum_{r \in \mathcal{S}_h} \delta_{h,r}(X) \right) \right| \\
&\leqslant \Lambda^{(t)} + (C-1)\Gamma^{(t)} + \Xi^{(t)} \kappa(\boldsymbol{X}) \\
&\leqslant 2 \left(\Lambda^{(t)} + (C-1)\Gamma^{(t)} \right).
\end{aligned} \tag{187}$$

We combine (186) and (187) and obtain:

$$\left| \frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}} \right| \leqslant \Theta(D\alpha^{(t)}) \mathfrak{S}(-yF(\boldsymbol{X})) \left(D\alpha^{(t)} \left(\Lambda^{(t)} + (C-1)\Gamma^{(t)} \right) \right)^{p-1} \Xi^{(t)}. \tag{188}$$

5. $\ell, m \neq \ell(\boldsymbol{X})$ and $\delta_s = 0$ for all s and $\delta_j = 0$. We first bound the outside sum.

$$\Xi^{(t)} \left| Cy + \sum_{h \neq \{\ell, \ell(\boldsymbol{X})\}} \sum_{r \in \mathcal{S}_h} \delta_r \right| \leqslant \Xi^{(t)} \kappa(\boldsymbol{X}). \tag{189}$$

We now bound the sum inside σ'. This sum is actually equal to the outside sum outside and we can therefore use the bound (189). Thus, the derivative is bounded as:

$$\left| \frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}} \right| \leqslant \Theta(D\alpha^{(t)}) \mathfrak{S}(-yF(\boldsymbol{X})) \left(D\alpha^{(t)} \kappa(\boldsymbol{X}) \Xi^{(t)} \right)^{p-1} (\Xi^{(t)})^2 \kappa(\boldsymbol{X}). \tag{190}$$

6. $\ell, m \neq \ell(\boldsymbol{X})$ and $\delta_s \neq 0$ for some s and $\delta_j = 0$. We first bound the outside sum. We apply Lemma G.3 and obtain:

$$\begin{aligned}
&\alpha^{(t)} \left| \Lambda^{(t)} \delta_i + \Gamma^{(t)} \sum_{r \in \mathcal{S}_\ell \setminus \{i\}} \delta_r + \Xi^{(t)} \left[Cy + \sum_{h \neq \{\ell, \ell(\boldsymbol{X})\}} \sum_{r \in \mathcal{S}_h} \delta_r \right] \right| \\
&\leqslant \alpha^{(t)} \left(\Lambda^{(t)} + (C-1)\Gamma^{(t)} + \Xi^{(t)} \left[C + \sum_{h \neq \{\ell, \ell(\boldsymbol{X})\}} \sum_{r \in \mathcal{S}_h} |\delta_r| \right] \right) \\
&\leqslant 2\alpha^{(t)} \left(\Lambda^{(t)} + (C-1)\Gamma^{(t)} \right).
\end{aligned} \tag{191}$$

We now bound the sum inside the power term. This sum is actually equal to the sum outside the power term and we can therefore use the bound (191). Thus, the derivative is bounded as:

$$\left| \frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}} \right| \leqslant \Theta(D\alpha^{(t)}) \mathfrak{S}(-yF(\boldsymbol{X})) \left(D\alpha^{(t)} (\Lambda^{(t)} + (C-1)\Gamma^{(t)}) \right)^{p-1} \Xi^{(t)} (\Lambda^{(t)} + (C-1)\Gamma^{(t)}). \tag{192}$$

7. $\ell, m \neq \ell(\boldsymbol{X})$ and $\delta_s = 0$ for all s and $\delta_j \neq 0$. We first bound the outside sum.

$$\begin{aligned}
&\Xi^{(t)} \left| \Lambda^{(t)} \delta_j + (C-1)\Gamma^{(t)} \delta_j + \Xi^{(t)} \left[C(\delta_j - y) + \sum_{h \neq \{\ell(X), \ell\}} \sum_{r \in \mathcal{S}_h} (\delta_j - \delta_r) \right] \right| \\
&\leqslant \Xi^{(t)} \left(\Xi^{(t)} \kappa(\boldsymbol{X}) + \mathbf{1}_{\delta_j \neq 0} \right).
\end{aligned} \tag{193}$$

We now bound the sum inside σ'.

$$\Xi^{(t)}\Big|Cy + \sum_{h\neq\{\ell(\boldsymbol{X}),\ell\}}\sum_{r\in\mathcal{S}_h}\delta_r\Big| \leqslant \Xi^{(t)}\kappa(\boldsymbol{X}). \tag{194}$$

Using (193) and (194), the bound on the derivative is:

$$\Big|\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\Big| \leqslant \Theta(D\alpha^{(t)})\mathfrak{S}(-yF(\boldsymbol{X}))\Big(D\alpha^{(t)}\Xi^{(t)}\kappa(\boldsymbol{X})\Big)^{p-1}\Xi^{(t)}\Big(\Xi^{(t)}\kappa(\boldsymbol{X}) + \mathbf{1}_{\delta_j\neq 0}\Big). \tag{195}$$

8. $\boldsymbol{\ell, m \neq \ell(X)}$ and $\boldsymbol{\delta_s(X) \neq 0}$ for some s and $\boldsymbol{\delta_j \neq 0}$: We first bound the outside sum. We apply Lemma G.4 to get:

$$\begin{aligned}
&\Xi^{(t)}\Big|\Lambda^{(t)}(\delta_j-\delta_i) + \Gamma^{(t)}\sum_{r\in\mathcal{S}_\ell\setminus\{i\}}(\delta_j-\delta_r) + \Xi^{(t)}\Big[C(\delta_j-y) + \sum_{h\neq\{\ell(\boldsymbol{X}),\ell\}}\sum_{r\in\mathcal{S}_h}(\delta_j-\delta_r)\Big]\Big| \\
&\leqslant \Xi^{(t)}\Big(\Lambda^{(t)} + (C-1)\Gamma^{(t)} + \Xi^{(t)}\kappa(\boldsymbol{X}) + \mathbf{1}_{\delta_j\neq 0}\Big) \\
&\leqslant \Theta(\Xi^{(t)}).
\end{aligned} \tag{196}$$

We now bound the sum inside the power term. We apply Lemma G.3 and get:

$$\begin{aligned}
\Big|\Lambda^{(t)}\delta_i + \Gamma^{(t)}\sum_{r\in\mathcal{S}_\ell\setminus\{i\}}\delta_r + \Xi^{(t)}\Big(Cy + \sum_{h\neq\{\ell(\boldsymbol{X}),\ell\}}\sum_{r\in\mathcal{S}_h}\delta_r\Big)\Big| &\leqslant \Lambda^{(t)} + (C-1)\Gamma^{(t)} + \Xi^{(t)}\kappa(\boldsymbol{X}) \\
&\leqslant \Theta\Big(\Lambda^{(t)} + (C-1)\Gamma^{(t)}\Big).
\end{aligned} \tag{197}$$

We plug (196) and (197) to obtain the derivative.

$$\Big|\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\Big| \leqslant \Theta(D\alpha^{(t)})\mathfrak{S}(-yF(\boldsymbol{X}))\Big(D\alpha^{(t)}\big(\Lambda^{(t)} + (C-1)\Gamma^{(t)}\big)\Big)^{p-1}\Xi^{(t)}. \tag{198}$$

Putting all the pieces together. We now bound the derivative of the population loss. Using Tower property and and Lemma I.1, we have:

$$\mathbb{E}_X\left[\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\right] = \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{\pi_1,\pi_2}\left[\frac{\partial L(\pi(\boldsymbol{X}))}{\partial A_{i,j}}\Big|\boldsymbol{X}\right]\right]. \tag{199}$$

For a *fixed* $\boldsymbol{X}$, we now bound the derivative of the loss evaluated in $\boldsymbol{X}$. For $i \in \mathcal{S}_\ell$ and $j \in \mathcal{S}_m$, we distinguish the eight possible events depending on the randomness of π.

- **Event a**: "$\ell = \ell(\pi(\boldsymbol{X}))$ and $\delta_j = 0$" occurs with probability $(1-q)/L$.
- **Event b**: "$\ell = \ell(\pi(\boldsymbol{X}))$ and $\delta_j \neq 0$" occurs with probability q/L.
- **Event c**: "$\ell \neq \ell(\pi(\boldsymbol{X}))$ and $\delta_s = 0$ for all s and $m = \ell(\pi(\boldsymbol{X}))$" occurs with probability $(1-q)^C/L$.
- **Event d$_k$**: "$\ell \neq \ell(\pi(\boldsymbol{X}))$ and $\#\{s : \delta_s \neq 0\} = k$ for $1 \leqslant k \leqslant C$ and $m = \ell(\pi(\boldsymbol{X}))$" occurs with probability $\binom{C}{k}q^k(1-q)^{C-k}/L$.
- **Event e**: "$\ell, m \neq \ell(\pi(\boldsymbol{X}))$ and $\delta_j = 0$ and $\delta_s = 0$ for all s" occurs with probability $(1-2/L)(1-q)^{C+1}$.
- **Event f$_k$**: "$\ell, m \neq \ell(\pi(\boldsymbol{X}))$ and $\#\{s : \delta_s \neq 0\} = k$ for $1 \leqslant k \leqslant C$ and $\delta_j = 0$" occurs with probability $(1-2/L)(1-q)\binom{C}{k}q^k(1-q)^{C-k}$.
- **Event g**: "$\ell, m \neq \ell(\pi(\boldsymbol{X}))$ and $\delta_s = 0$ for all s and $\delta_j \neq 0$" occurs with probability $(1-2/L)(1-q)^C q$.
- **Event h$_k$**: "$\ell, m \neq \ell(\pi(\boldsymbol{X}))$ and $\#\{s : \delta_s \neq 0\} = k$ for $0 \leqslant k \leqslant C$ and $\delta_j \neq 0$" occurs with probability $(1-2/L)q\binom{C}{k}q^k(1-q)^{C-k}$

Since events a and e are the ones with highest probabilities, the derivative of the loss is bounded by the expectations conditioned on these events. We have:

$$\begin{aligned}
&\mathbb{E}_{\pi_1,\pi_2}\Big[\Big|\frac{\partial L(\pi(\boldsymbol{X}))}{\partial A_{i,j}}\Big|\Big] \\
&\leqslant \frac{\Theta(D\alpha^{(t)})}{L}\Xi^{(t)}\mathfrak{S}(-yF(\boldsymbol{X}))\big(D\alpha^{(t)}(\Lambda^{(t)}+(C-1)\Gamma^{(t)})\big)^{p-1}(\Lambda^{(t)}+(C-1)\Gamma^{(t)}) \\
&+\Theta(D\alpha^{(t)})\Xi^{(t)}\mathfrak{S}(-yF(\boldsymbol{X}))\mathbb{E}_{\pi_1,\pi_2}\Big[\big(D\alpha^{(t)}\kappa(\pi(\boldsymbol{X}))\Xi^{(t)}\big)^{p-1}\Xi^{(t)}\kappa(\pi(\boldsymbol{X}))\Big|\mathbf{e}\Big].
\end{aligned} \tag{200}$$

We now apply Induction Hypothesis B.1 and Lemma K.3 and finally obtain:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}}\Big[\Big|\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\Big|\Big] &\leqslant \Theta(C^2\alpha^{(t)})\mathbb{E}_{\boldsymbol{X}}[\mathfrak{S}(-yF(\boldsymbol{X}))]\big(D\alpha^{(t)}(\Lambda^{(t)}+(C-1)\Gamma^{(t)})\big)^{p-1}\Xi^{(t)}\frac{\lambda_0}{D} \\
&+\Theta(\alpha^{(t)})\mathbb{E}_{\boldsymbol{X}}[\mathfrak{S}(-yF(\boldsymbol{X}))]\big(D\alpha^{(t)}qD\log(d)\Xi^{(t)}\big)^{p-1}\Xi^{(t)}qD\log(d).
\end{aligned} \tag{201}$$

Since $\mathbb{E}_{\boldsymbol{X}}[\mathfrak{S}(-yF(\boldsymbol{X}))] \leqslant 1$, we thus plug (201) in the update of $\rho^{(t)}$ to obtain the aimed result.

□

Corollary G.2. *Let $T > 0$ be the time where the population loss is $o(1)$ and $t \leqslant T$. The update of $\rho^{(t)}$ satisfies:*

$$|\rho^{(t+1)}| \leqslant |\rho^{(t)}| + \eta \cdot \mathrm{polylog}(d)(\alpha^{(t)})^p\left(\frac{1}{D}+\frac{\lambda_0}{D}\Xi^{(t)}(G^{(t)})^{p-1}\right).$$

Proof of Corollary G.2. Using Lemma G.2, $C = \mathrm{polylog}(d)$ and $qD \leqslant 1$, $\rho^{(t)}$'s update is:

$$|\rho^{(t+1)}| \leqslant |\rho^{(t)}| + \eta(\alpha^{(t)})^p\mathrm{polylog}(d)\left((D\Xi^{(t)})^{p-1}\Xi^{(t)}+\frac{\lambda_0}{D}(G^{(t)})^{p-1}\Xi^{(t)}\right). \tag{202}$$

Lastly, we apply Induction Hypothesis B.1 to replace $\Xi^{(t)}$ by its value in (202) and thus obtain the aimed result. □

G.3 Auxiliary lemmas

Lemma G.3. *Let $t > 0$. In the idealized learning process, with high probability, we have:*

$$\Lambda^{(t)} + (C-1)\Gamma^{(t)} + y\Xi^{(t)}\sum_{\ell\neq\ell(\boldsymbol{X})}\sum_{r\in\mathcal{S}_\ell}\delta_r = \Theta\big(\Lambda^{(t)}+(C-1)\Gamma^{(t)}\big).$$

Proof of Lemma G.3. We first bound the sum with factor $\Xi^{(t)}$. Using Induction Hypothesis B.1 and Lemma K.3, we have:

$$\Big|y\Xi^{(t)}\sum_{\ell\neq\ell(\boldsymbol{X})}\sum_{r\in\mathcal{S}_\ell}\delta_r\Big| \leqslant \Theta(q)\log(d) = \frac{1}{D}\mathrm{poly}(C)\log(d). \tag{203}$$

We now bound $\Lambda^{(t)} + (C-1)\Gamma^{(t)}$. Using Induction Hypothesis B.1, we have:

$$\frac{1}{D}\left(e^{\mathrm{polyloglog}(d)}+\Omega(C)\right) \leqslant (\Lambda^{(t)}+(C-1)\Gamma^{(t)}) \leqslant \frac{1}{D}\left(e^{\mathrm{polyloglog}(d)}+\lambda_0\right). \tag{204}$$

Since $\mathrm{poly}(C)\log(d) \leqslant e^{\mathrm{polyloglog}(d)}/2$, we combine (203) and (204) to get the aimed result. □

Lemma G.4. *Let $t > 0$. In the idealized learning process, we have with high probability:*

$$\Lambda^{(t)} + (C-1)\Gamma^{(t)} + \Xi^{(t)}\sum_{h\neq\ell(\boldsymbol{X})}\sum_{r\in\mathcal{S}_h}|\delta_r| \leqslant \Theta(1).$$

Proof of Lemma G.4. We successively apply Lemma K.3 and Induction Hypothesis B.1 to get the desired bound. Indeed, we have:

$$\begin{aligned}\Lambda^{(t)} + (C-1)\Gamma^{(t)} + \Xi^{(t)} \sum_{h \neq \ell(\boldsymbol{X})} \sum_{r \in \mathcal{S}_h} |\delta_r| &\leqslant \Lambda^{(t)} + (C-1)\Gamma^{(t)} + \Xi^{(t)} qD\log(d) \\ &\leqslant \frac{1}{D}\left(e^{\mathrm{polyloglog}(d)} + \lambda_0 + qD\log(d)\right) \\ &= \Theta(1).\end{aligned}$$

□

H Gradients

In this section, we present the gradients of the loss $\mathcal{L}$ with respect to $\boldsymbol{v}$ and $\boldsymbol{A}_{i,j}$.

Lemma H.1. *Let $(\boldsymbol{X}, y)$ be a data-point. Then, the gradient of $L(\boldsymbol{X})$ with respect to $\boldsymbol{v}$ is:*

$$-\nabla_v L(\boldsymbol{X}) = Dy\mathfrak{S}(-yF(\boldsymbol{X})) \sum_{i=1}^{D} \sigma'\left(D \sum_{r \in [D]} \boldsymbol{S}_{i,r} \langle \boldsymbol{v}, \boldsymbol{X}_r \rangle\right) \sum_{j=1}^{D} \boldsymbol{S}_{i,j} \boldsymbol{X}_j.$$

Lemma H.2. *Let $(\boldsymbol{X}, y)$ be a data-point and $i, j \in [D]$. The derivative of $L(\boldsymbol{X})$ with respect to $A_{i,j}$ is:*

$$-\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}} = pDy\mathfrak{S}(-yF(\boldsymbol{X}))\left(D \sum_{r \in [D]} \boldsymbol{S}_{i,r} \langle \boldsymbol{v}, \boldsymbol{X}_r \rangle\right)^{p-1} \sum_{r \neq j} \boldsymbol{S}_{i,r} \langle \boldsymbol{v}, \boldsymbol{X}_j - \boldsymbol{X}_r \rangle.$$

I Invariance of the problem

I.1 Invariance of the parameters

Lemma 4.2. *For all $t \in [T]$, there exists $\alpha^{(t)} \in \mathbb{R}$ such that $\boldsymbol{v}^{(t)} = \alpha^{(t)} \boldsymbol{w}^*$.*

Proof of Lemma 4.2. The proof is by induction. Our induction hypothesis is $\boldsymbol{v}^{(t)} = \alpha^{(t)} \boldsymbol{w}^*$ for all $t \geqslant 0$. For $t = 0$, we know that $\boldsymbol{v}^{(0)} = \alpha^{(0)} \boldsymbol{w}^* \in \mathrm{span}(\boldsymbol{w}^*)$. Assume that $\boldsymbol{v}^{(t)} = \alpha^{(t)} \boldsymbol{w}^*$. Let's show that there exists $\alpha^{(t+1)} \in \mathbb{R}$ such that $\boldsymbol{v}^{(t+1)} = \alpha^{(t+1)} \boldsymbol{w}^*$. From the update rule, we have:

$$\begin{aligned}\boldsymbol{v}^{(t+1)} &= \boldsymbol{v}^{(t)} + \eta \mathbb{E}\Big[y\mathfrak{S}(-yF(\boldsymbol{X})) \sum_{i \in [D]} \sigma'(\langle \boldsymbol{v}^{(t)}, \boldsymbol{O}_i^{(t)} \rangle) \sum_{r \in [D]} S_{i,r} \boldsymbol{X}_r\Big] \\ &= \boldsymbol{v}^{(t)} + \eta \mathbb{E}\Big[y\mathfrak{S}(-yF(\boldsymbol{X})) \sum_{i \in [D]} \sigma'(\langle \boldsymbol{v}^{(t)}, \boldsymbol{O}_i^{(t)} \rangle) \sum_{r \in [D]} S_{i,r} \mathbb{E}_{\boldsymbol{\xi}|_{\mathrm{span}(\boldsymbol{w}^*)^\perp}}[\boldsymbol{X}_r]\Big], \qquad (205)\end{aligned}$$

where $\mathbb{E}_{\boldsymbol{\xi}|_{\mathrm{span}(\boldsymbol{w}^*)^\perp}}$ is the expectation with respect to the projection of the noise vectors $\boldsymbol{\xi}_r$ onto $\mathrm{span}(\boldsymbol{w}^*)^\perp$. Using the definition of the data distribution and $\mathbb{E}[\boldsymbol{\xi}_r] = 0$, we simplify the update (205):

$$\begin{aligned}\boldsymbol{v}^{(t+1)} = \boldsymbol{v}^{(t)} &+ \eta \mathbb{E}\Big[\mathfrak{S}(-yF(\boldsymbol{X})) \sum_{i \in [D]} \sigma'(\langle \boldsymbol{v}^{(t)}, \boldsymbol{O}_i^{(t)} \rangle) \sum_{r \in \mathcal{S}_{\ell(X)}} S_{i,r}\Big] \boldsymbol{w}^* \\ &+ \eta \mathbb{E}\Big[y\mathfrak{S}(-yF(\boldsymbol{X})) \sum_{i \in [D]} \sigma'(\langle \boldsymbol{v}^{(t)}, \boldsymbol{O}_i^{(t)} \rangle) \sum_{h \neq \ell(X)} \sum_{r \in \mathcal{S}_h} S_{i,r} \delta_r\Big] \boldsymbol{w}^*. \qquad (206)\end{aligned}$$

Since $\boldsymbol{v}^{(t)} = \alpha^{(t)} \boldsymbol{w}^*$, there exists $\alpha^{(t+1)} \in \mathbb{R}$ such that $\boldsymbol{v}^{(t+1)} = \alpha^{(t+1)} \boldsymbol{w}^*$.

□

Lemma 4.1. *There exist $\beta = \sigma_{\boldsymbol{A}}$, $\gamma^{(t)}, \rho^{(t)} \in \mathbb{R}$ such that for all $t \geqslant 0$:*

1. *for all $i \in [D]$, $A_{i,i}^{(t)} = \beta$.*

2. *for all* $i, j \in [D]$ *such that* $i, j \in \mathcal{S}_\ell$ *for some* $\ell \in [L]$, $A_{i,j}^{(t)} = \gamma^{(t)}$.

3. *for all* $i, j \in [D]$ *such that* $i \in \mathcal{S}_\ell$ *and* $j \in \mathcal{S}_m$ *for some* $\ell, m \in [L]$ *with* $\ell \neq m$, $A_{i,j}^{(t)} = \rho^{(t)}$.

Proof of Lemma 4.1. We initialize $A_{i,i}^{(0)} = \sigma_{\boldsymbol{A}}$ and do not update $A_{i,i}$, Thus, for all $t \geqslant 0$, we have $A_{i,i}^{(t)} = \sigma_{\boldsymbol{A}}$.

The remaining of the proof is by induction. For $i, j \in \mathcal{S}_\ell$, we initialize $A_{i,j}^{(0)} = 0 = \gamma^{(0)}$. For $i \in \mathcal{S}_\ell$ and $j \in \mathcal{S}_m$ with $\ell \neq m$, $A_{i,j}^{(0)} = 0 = \rho^{(0)}$. The induction hypothesis is true for $t = 0$.

We assume that for all $i, j \in \mathcal{S}_\ell$, $A_{i,j}^{(t)} = \gamma^{(t)}$ and for $i \in \mathcal{S}_\ell$ and $j \in \mathcal{S}_m$, $A_{i,j}^{(t)} = \rho^{(t)}$. Let's first prove that $A_{i,j}^{(t+1)} = A_{k,n}^{(t+1)}$ for $i, j \in \mathcal{S}_\ell$ and $k, n \in \mathcal{S}_{\ell'}$. Since the GD update is $A_{i,j}^{(t+1)} = A_{i,j}^{(t)} - \eta \mathbb{E}_{\boldsymbol{X}}\left[\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\right]$, it's sufficient to prove that $\mathbb{E}_{\boldsymbol{X}}\left[\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\right] = \mathbb{E}_{\boldsymbol{X}}\left[\frac{\partial L(\boldsymbol{X})}{\partial A_{k,n}}\right]$. Let $\pi_1 \colon [L] \to [L]$, $\pi_2 \colon [C] \to [C]$ and $\pi = (\pi_1, \pi_2)$ be permutations. From Lemma I.1, we know that $\boldsymbol{X}$ and $\pi(\boldsymbol{X})$ have the same distribution which implies $\mathbb{E}_{\boldsymbol{X}}\left[\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\right] = \mathbb{E}_{\pi(\boldsymbol{X})}\left[\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\right]$. Therefore, we have:

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}\left[\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\right] \\
=&\mathbb{E}_{\boldsymbol{X}}\Bigg[y\mathfrak{S}(-yF(\pi(\boldsymbol{X})))\sigma'\Bigg(\sum_{h=1}^{L}\sum_{r\in\mathcal{S}_h}\frac{e^{\langle \boldsymbol{A}^{(t)}\boldsymbol{p}_i,\, \boldsymbol{p}_r\rangle}}{\sum_{s=1}^{D} e^{\langle \boldsymbol{A}^{(t)}\boldsymbol{p}_i,\, \boldsymbol{p}_s\rangle}}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi_1(h),\pi_2(r)}\rangle\Bigg)\cdot \\
&\frac{e^{\langle \boldsymbol{A}^{(t)}\boldsymbol{p}_i,\, \boldsymbol{p}_j\rangle}}{\sum_{s=1}^{D} e^{\langle \boldsymbol{A}^{(t)}\boldsymbol{p}_i,\, \boldsymbol{p}_s\rangle}}\sum_{k=1}^{L}\sum_{m\in\mathcal{S}_k}\frac{e^{\langle \boldsymbol{A}^{(t)}\boldsymbol{p}_i,\, \boldsymbol{p}_m\rangle}}{\sum_{s=1}^{D} e^{\langle \boldsymbol{A}^{(t)}\boldsymbol{p}_i,\, \boldsymbol{p}_s\rangle}}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi_1(k),\pi_2(m)} - \boldsymbol{X}_{\pi_1(\ell),\pi_2(j)}\rangle\Bigg]
\end{aligned}
\tag{207}
$$

Using Induction Hypothesis B.1, we simplify (207) as

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}\left[\frac{\partial \mathcal{L}^{(t)}(\boldsymbol{X})}{\partial A_{i,j}}\right] \\
=&\mathbb{E}_{\boldsymbol{X}}\Bigg[y\mathfrak{S}(-yF(\pi(\boldsymbol{X})))\sigma'\Bigg(\Lambda^{(t)}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi_1(\ell),\pi_2(i)}\rangle + \Gamma^{(t)}\sum_{m\in\mathcal{S}_\ell\setminus\{i\}}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi_1(\ell),\pi_2(m)}\rangle \\
&+ \Xi^{(t)}\sum_{r\notin\mathcal{S}_\ell}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi(r)}\rangle\Bigg)\Gamma^{(t)}\Bigg(\Lambda^{(t)}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi_1(\ell),\pi_2(i)} - \boldsymbol{X}_{\pi_1(\ell),\pi_2(j)}\rangle \\
&+ \Gamma^{(t)}\sum_{m\in\mathcal{S}_\ell\setminus\{i\}}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi_1(\ell),\pi_2(m)} - \boldsymbol{X}_{\pi_1(\ell),\pi_2(j)}\rangle + \Xi^{(t)}\sum_{r\notin\mathcal{S}_\ell}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi(r)} - \boldsymbol{X}_{\pi_1(\ell),\pi_2(j)}\rangle\Bigg)\Bigg].
\end{aligned}
\tag{208}
$$

We now set π_1 and π_2 such that $\pi_1(\ell) = \ell'$, $\pi_2(i) = k$ and $\pi_2(j) = n$. Using this choice along with $F(\boldsymbol{X}) = F(\pi(\boldsymbol{X}))$ (Lemma I.1), we finally have in (208)

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}\left[\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\right] \\
=&\mathbb{E}_{\boldsymbol{X}}\Bigg[y\mathfrak{S}(-yF(\pi(\boldsymbol{X})))\sigma'\Bigg(\Lambda^{(t)}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\ell',k}\rangle + \Gamma^{(t)}\sum_{m\in\mathcal{S}_{\ell'}\setminus\{i\}}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\ell',m}\rangle \\
&+ \Xi^{(t)}\sum_{r\notin\mathcal{S}_{\ell'}}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{r}\rangle\Bigg)\Gamma^{(t)}\Bigg(\Lambda^{(t)}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\ell',k} - \boldsymbol{X}_{\ell',n}\rangle \\
&+ \Gamma^{(t)}\sum_{m\in\mathcal{S}_{\ell'}\setminus\{k\}}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\ell',m} - \boldsymbol{X}_{\ell',n}\rangle + \Xi^{(t)}\sum_{r\notin\mathcal{S}_\ell}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{r} - \boldsymbol{X}_{\ell',n}\rangle\Bigg)\Bigg] \\
=&\mathbb{E}_{\boldsymbol{X}}\left[\frac{\partial \mathcal{L}^{(t)}(\boldsymbol{X})}{\partial A_{k,n}}\right].
\end{aligned}
\tag{209}
$$

Therefore, (209) implies that $A_{i,j}^{(t+1)} = A_{k,n}^{(t+1)}$ thus proving the induction hypothesis. Let's now show that for $i \in \mathcal{S}_\ell, j \in \mathcal{S}_m$ and $k \in \mathcal{S}_{\ell'}, j \in \mathcal{S}_{m'}$, we have $A_{i,j}^{(t+1)} = A_{k,n}^{(t+1)}$. We apply a similar argument as above. Using Induction Hypothesis B.1, we have:

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}\left[\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\right] \\
=&\mathbb{E}_{\boldsymbol{X}}\Bigg[y\mathfrak{S}(-yF(\pi(\boldsymbol{X})))\sigma'\Bigg(\Lambda^{(t)}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi_1(\ell),\pi_2(i)}\rangle + \Gamma^{(t)}\sum_{r\in\mathcal{S}_\ell\setminus\{i\}}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi_1(\ell),\pi_2(r)}\rangle \\
&+ \Xi^{(t)}\sum_{r\notin\mathcal{S}_\ell}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi(r)}\rangle\Bigg)\Xi^{(t)}\Bigg(\Lambda^{(t)}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi_1(\ell),\pi_2(i)} - \boldsymbol{X}_{\pi_1(m),\pi_2(j)}\rangle \\
&+ \Gamma^{(t)}\sum_{r\in\mathcal{S}_\ell\setminus\{i\}}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi_1(\ell),\pi_2(r)} - \boldsymbol{X}_{\pi_1(m),\pi_2(j)}\rangle + \Xi^{(t)}\sum_{r\notin\mathcal{S}_\ell}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi(r)} - \boldsymbol{X}_{\pi_1(m),\pi_2(j)}\rangle\Bigg)\Bigg].
\end{aligned}
\tag{210}
$$

We now set π_1 and π_2 such that $\pi_1(\ell) = \ell'$, $\pi_1(m) = m'$, $\pi_2(i) = k$ and $\pi_2(j) = n$. Using this choice, we finally have in (210)

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}\left[\frac{\partial L(\boldsymbol{X})}{\partial A_{i,j}}\right] \\
=&\mathbb{E}_{\boldsymbol{X}}\Bigg[y\mathfrak{S}(-yF(\pi(\boldsymbol{X})))\sigma'\Bigg(\Lambda^{(t)}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\ell',k}\rangle + \Gamma^{(t)}\sum_{r\in\mathcal{S}_{\ell'}\setminus\{k\}}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\ell',r}\rangle \\
&+ \Xi^{(t)}\sum_{r\notin\mathcal{S}_{\ell'}}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{r}\rangle\Bigg)\Xi^{(t)}\Bigg(\Lambda^{(t)}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\ell',k} - \boldsymbol{X}_{m',n}\rangle \\
&+ \Gamma^{(t)}\sum_{r\in\mathcal{S}_{\ell'}\setminus\{k\}}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\ell',r} - \boldsymbol{X}_{m',n}\rangle + \Xi^{(t)}\sum_{r\notin\mathcal{S}_\ell}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{r} - \boldsymbol{X}_{m',n}\rangle\Bigg)\Bigg] \\
=&\mathbb{E}_{\boldsymbol{X}}\left[\frac{\partial L(\boldsymbol{X})}{\partial A_{k,n}}\right].
\end{aligned}
\tag{211}
$$

Therefore, (211) implies that $A_{i,j}^{(t+1)} = A_{k,n}^{(t+1)}$ thus proving the induction hypothesis.

□

I.2 Invariance by permutation

Lemma I.1. *Let $\pi_1 \colon [L] \to [L]$ and $\pi_2 \colon [C] \to [C]$ be two permutations and $\pi = (\pi_1, \pi_2)$. Let $(\boldsymbol{X}, \cdot) \sim \mathcal{D}$. Then, we have:*

1. *permutation-invariant distribution: $\boldsymbol{X}$ has the same distribution as $\pi(\boldsymbol{X})$.*
2. *permutation-invariant model: $F(\boldsymbol{X}) = F(\pi(\boldsymbol{X}))$.*

Proof of Lemma I.1. To show that $\boldsymbol{X}$ and $\pi(\boldsymbol{X})$ have to same distribution, it is sufficient to show that the items 1 to 6 hold in our definition of the data distribution. We still have that the label y is uniformly sampled on $\{-1, 1\}$. Let $\pi(\boldsymbol{X}) = (\overline{\boldsymbol{X}}_1, \ldots, \overline{\boldsymbol{X}}_D)$ where $\overline{\boldsymbol{X}}_i$ for some j. For 2, the number of tokens is still D after permutation π. For 4, we define the same partition with $\bar{\mathcal{S}}_l = \bar{\mathcal{S}}_{\pi_1(l)}$. Besides, we have $\ell(\overline{\boldsymbol{X}}) = \pi_1(\ell(\boldsymbol{X}))$ that is also uniformly sampled on $[L]$ since the permutation on a uniform distribution is also uniform. For $i \in \mathcal{S}_{\ell(\overline{\boldsymbol{X}})}$, we have that $\overline{\boldsymbol{X}}_i$ writes $\boldsymbol{X}_{\pi_2(k)}$ for some $k \in \mathcal{S}_{\ell(\boldsymbol{X})}$ so that we do have $\overline{\boldsymbol{X}}_i = y\boldsymbol{w}^* + \boldsymbol{\xi}_i$. The same goes for 6 when $\ell \neq \ell(\overline{\boldsymbol{X}})$: with $\ell \neq \ell(\overline{\boldsymbol{X}})$, $\overline{\boldsymbol{X}}_j = \delta_{\ell,j}\boldsymbol{w}^* + \boldsymbol{\xi}_j$, where $\delta_j = 1$ with probability $q/2$, -1 with the same probability and 0 otherwise.

Let $\boldsymbol{X}$ be a data-point. Using Lemma 4.1, we rewrite $F(\boldsymbol{X})$ as

$$
F(\boldsymbol{X}) = \sum_{m=1}^{D}\sigma\Big(\Lambda^{(t)}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_m\rangle + \Gamma^{(t)}\sum_{r\in\mathcal{S}_\ell\setminus\{m\}}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_r\rangle + \Xi^{(t)}\sum_{h\in[L]\setminus\{\ell\}}\sum_{j\in\mathcal{S}_h}\langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_j\rangle\Big). \tag{212}
$$

Let $\pi = (\pi_1, \pi_2)$ be a permutation. For a given $\ell \in [L]$ and $m \in \mathcal{S}_\ell$, assume that $\pi_1(\ell') = \ell$ and $\pi_2(m') = m$. Using Lemma 4.1, we have:

$$\begin{aligned}
F(\pi(\boldsymbol{X})) &= \sum_{\ell=1}^{L} \sum_{m \in \mathcal{S}_\ell} \sigma\Big(\Lambda^{(t)} \langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi_2(m')} \rangle \\
&\quad + \Gamma^{(t)} \sum_{r \in \mathcal{S}_{\pi_1(\ell')} \setminus \{\pi_2(m')\}} \langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi_2(r)} \rangle + \Xi^{(t)} \sum_{h' \in [L] \setminus \{\ell'\}} \sum_{j \in \mathcal{S}_{h'}} \langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{\pi_2(j')} \rangle \Big) \\
&= \sum_{\ell=1}^{L} \sum_{m \in \mathcal{S}_\ell} \sigma\Big(\Lambda^{(t)} \langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{m} \rangle \\
&\quad + \Gamma^{(t)} \sum_{r'' \in \mathcal{S}_\ell \setminus \{m\}} \langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{r''} \rangle + \Xi^{(t)} \sum_{j'' \notin \mathcal{S}_\ell} \langle \boldsymbol{v}^{(t)}, \boldsymbol{X}_{j''} \rangle \Big) \\
&= F(\boldsymbol{X}).
\end{aligned}$$

□

J Justification of our data distribution

In this section, we justify why the distribution $\mathcal{D}$ (Definition 3.1) is relevant. We first show that linear classifiers poorly generalize (subsection J.1). We then show that there exists classifiers that generalize without learning patch association (subsection J.2).

J.1 Generalized linear models poorly generalize

Theorem 3.1. *Let $\mathcal{D}$ be as defined in Definition 3.1. Let $g(\boldsymbol{X}) = \phi\left(\sum_{j=1}^{D} \langle \boldsymbol{w}_j, \boldsymbol{X}_j \rangle\right)$ be a generalized linear model. Then, g does not fit the labeling function i.e. $\mathbb{P}[f^*(\boldsymbol{X}) g(\boldsymbol{X}) \leqslant 0] \geqslant 1/8$.*

Proof of Theorem 3.1. For every data point $\boldsymbol{X}$, consider $\Delta(\boldsymbol{X}) := \sum_{j \in [D]} \delta_j$, it is very easy to see that for every integer p, as long as $\Pr[\Delta = p] = \Omega(1/\text{polylog}(d))$, we have that:

$$\frac{\Pr[\Delta = p]}{\Pr[\Delta = p - 2C]} = 1 - o(1) \tag{213}$$

Consider two independently sampled data points, $\boldsymbol{X}, \boldsymbol{X}'$ with label $1, -1$ respectively, consider the event when $\Delta(\boldsymbol{X}) = p - 2C, \Delta(\boldsymbol{X}') = p$ and all the noises ξ_i, ξ_i' of $\boldsymbol{X}, \boldsymbol{X}'$ satisfies $\xi_i = \xi_i'$, then we know that

$$\sum_{j \in [D]} \langle w_j, \boldsymbol{X}_j \rangle = \sum_{j \in [D]} \langle w_j, \boldsymbol{X}_j' \rangle \tag{214}$$

By Eq (213) we also know that the density of $\boldsymbol{X}$ and $\boldsymbol{X}'$ under the data-generation distribution satisfies

$$p(\boldsymbol{X}) = (1 \pm o(1)) p(\boldsymbol{X}')$$

Now, we know that

$$P_0 := \mathbb{P}[f^*(\boldsymbol{X})g(\boldsymbol{X}) \leqslant 0 \mid y(\boldsymbol{X}) = 1] = 2\int_{\boldsymbol{X}} 1_{f^*(\boldsymbol{X})g(\boldsymbol{X})\leqslant 0} p(\boldsymbol{X}, y(\boldsymbol{X}) = 1) d\boldsymbol{X} \tag{215}$$

$$= 2\int_{\boldsymbol{X}} 1_{f^*(\boldsymbol{X})g(\boldsymbol{X})\leqslant 0} p(\boldsymbol{X}', y(\boldsymbol{X}') = -1) d\boldsymbol{X} \pm o(1) \tag{216}$$

$$= 2\int_{\boldsymbol{X}} 1_{f^*(\boldsymbol{X})g(\boldsymbol{X}')\leqslant 0} p(\boldsymbol{X}', y(\boldsymbol{X}') = -1) d\boldsymbol{X} \pm o(1) \tag{217}$$

$$= 2\int_{\boldsymbol{X}'} 1_{f^*(\boldsymbol{X}')g(\boldsymbol{X}')\geqslant 0} p(\boldsymbol{X}', y(\boldsymbol{X}') = -1) d\boldsymbol{X}' \pm o(1) \tag{218}$$

$$= 1 - \Pr[f^*(\boldsymbol{X}')g(\boldsymbol{X}') < 0 \mid y(\boldsymbol{X}') = -1] \pm o(1) \tag{219}$$

$$\geqslant 1 - \Pr[f^*(\boldsymbol{X}')g(\boldsymbol{X}') \leqslant 0 \mid y(\boldsymbol{X}') = -1] \pm o(1) \tag{220}$$

$$= 1 - \Pr[f^*(\boldsymbol{X})g(\boldsymbol{X}) \leqslant 0 \mid y(\boldsymbol{X}) = -1] \pm o(1) \tag{221}$$

Therefore, $\mathbb{P}[f^*(\boldsymbol{X})g(\boldsymbol{X}) \leqslant 0 \mid y(\boldsymbol{X}) = 1] \geqslant \frac{1}{2} - o(1)$. □

J.2 Classifiers fitting the labelling function without patch association

Theorem 3.2. *Let $\mathcal{D}$ be defined as in Definition 3.1. There exists a (one-layer) transformer $\mathcal{M}$ so that $\mathbb{P}[f^*(\boldsymbol{X})\mathcal{M}(\boldsymbol{X}) \leqslant 0] = d^{-\omega(1)}$ but for all $\ell \in [L]$, $i \in \mathcal{S}_\ell$, $\mathrm{Top}_C\ \{\langle \boldsymbol{p}_i^{(\mathcal{M})}, \boldsymbol{p}_j^{(\mathcal{M})}\rangle\}_{j=1}^D \cap \mathcal{S}_\ell = \varnothing$.*

Proof of Theorem 3.2. We can consider a transformer in our setting, whose weights are defined as: $\boldsymbol{v} = \boldsymbol{w}^\star$, $A_{i,j} = \beta > 0$ for $i, j \in \mathcal{S}_\ell$. $\boldsymbol{A}_{i,j} = 2\beta$ for $i \in \mathcal{S}_\ell, j \in \mathcal{S}_{\ell+1}$ (We denote $\mathcal{S}_{L+1} = \mathcal{S}_1$). For a sufficiently large β, it is easy to check that $\mathbb{P}[f^*(\boldsymbol{X})\mathcal{M}(\boldsymbol{X}) \leqslant 0] = d^{-\omega(1)}$ but for all $\ell \in [L]$, $i \in \mathcal{S}_\ell$, $\mathrm{Top}_C\ \{\langle \boldsymbol{p}_i^{(\mathcal{M})}, \boldsymbol{p}_j^{(\mathcal{M})}\rangle\}_{j=1}^D = \mathcal{S}_{\ell+1} \cap \mathcal{S}_\ell = \varnothing$.

□

K Technical lemmas

In this section, we present the technical lemmas used in the paper.

K.1 Tensor Power Method

Lemma K.1. *Let $\{z^{(t)}\}_{t\geqslant 0}$ be a positive sequence defined by the following recursions*

$$\begin{cases} z^{(t+1)} \geqslant z^{(t)} + m(z^{(t)})^k \\ z^{(t+1)} \leqslant z^{(t)} + M(z^{(t)})^k \end{cases},$$

where $z^{(0)} > 0$ is the initialization, $k > 1$ is an integer and $m, M > 0$. Let $\upsilon > 0$ such that $z^{(0)} \leqslant \upsilon$. Then, the time $\mathcal{T}$ such that $z^{(t)} \geqslant \upsilon$ for all $t \geqslant \mathcal{T}$ is:

$$\mathcal{T} = \frac{3}{m(z^{(0)})^{k-1}} + \frac{2^{k+1}M}{m}\left\lceil \frac{\log(\upsilon/z^{(0)})}{\log(2)} \right\rceil.$$

Proof of Lemma K.1. Let $n \in \mathbb{N}^*$. Let T_n be the time where $z^{(t)} \geqslant 2^n z^{(0)}$. This time exists because $z^{(t)}$ is a non-decreasing sequence. We want to find an upper bound on this time. We start with the case $n = 1$. By summing the recursion, we have:

$$z^{(T_1)} \geqslant z^{(0)} + m\sum_{s=0}^{T_1-1}(z^{(s)})^k. \tag{222}$$

We use the fact that $z^{(s)} \geqslant z^{(0)}$ in (222) and obtain:

$$T_1 \leqslant \frac{z^{(T_1)} - z^{(0)}}{m(z^{(0)})^k}. \tag{223}$$

Now, we want to bound $z^{(T_1)} - z^{(0)}$. Using again the recursion and $z^{(T_1-1)} \leqslant 2z^{(0)}$, we have:

$$z^{(T_1)} \leqslant z^{(T_1-1)} + M(z^{(T_1-1)})^k \leqslant 2z^{(0)} + 2^k M(z^{(0)})^k. \tag{224}$$

Combining (223) and (224), we get a bound on T_1.

$$T_1 \leqslant \frac{1}{m(z^{(0)})^{k-1}} + \frac{2^k M}{m}. \tag{225}$$

Now, let's find a bound for T_n. Starting from the recursion and using the fact that $z^{(s)} \geqslant 2^{n-1}z^{(0)}$ for $s \geqslant T_{n-1}$ we have:

$$z^{(T_n)} \geqslant z^{(T_{n-1})} + m \sum_{s=T_{n-1}}^{T_n-1} (z^{(s)})^k \geqslant z^{(T_{n-1})} + 2^{k(n-1)} m (z^{(0)})^k (T_n - T_{n-1}). \tag{226}$$

On the other hand, by using $z^{(T_n-1)} \leqslant 2^n z^{(0)}$ we upper bound $z^{(T_n)}$ as follows.

$$z^{(T_n)} \leqslant z^{(T_n-1)} + M(z^{(T_n-1)})^k \leqslant 2^n z^{(0)} + 2^{kn} M (z^{(0)})^k. \tag{227}$$

Besides, we know that $z^{(T_{n-1})} \geqslant 2^{n-1}z^{(0)}$. Therefore, we upper bound $z^{(T_n)} - z^{(T_{n-1})}$ as

$$z^{(T_n)} - z^{(T_{n-1})} \leqslant 2^{n-1} z^{(0)} + 2^{kn} M (z^{(0)})^k. \tag{228}$$

Combining (226) and (228) yields:

$$T_n \leqslant T_{n-1} + \frac{1}{2^{(k-1)(n-1)} m (z^{(0)})^{k-1}} + \frac{2^k M}{m}. \tag{229}$$

We now sum (229) for $n = 2, \dots, n$, use (225) and obtain:

$$T_n \leqslant T_1 + \frac{2}{m(z^{(0)})^{k-1}} + \frac{2^k M n}{m} \leqslant \frac{3}{m(z^{(0)})^{k-1}} + \frac{2^k M(n+1)}{m} \leqslant \frac{3}{m(z^{(0)})^{k-1}} + \frac{2^{k+1} M n}{m}. \tag{230}$$

Lastly, we know that n satisfies $2^n z^{(0)} \geqslant \upsilon$ which implies $n = \left\lceil \frac{\log(\upsilon/z_0)}{\log(2)} \right\rceil$ in (230). □

Lemma K.2. *Let $\{z^{(t)}\}_{t \geqslant 0}$ be a positive sequence defined by the following recursions*

$$\begin{cases} z^{(t+1)} \geqslant z^{(t)} + m(z^{(t)})^k \\ z^{(t+1)} \leqslant z^{(t)} + M(z^{(t)})^k \end{cases},$$

where $z^{(0)} > 0$, $k > 1$ is an integer and $m, M > 0$. Let $\upsilon > 0$ such that $z^{(0)} \leqslant \upsilon$ and $\mathcal{T}$ be the time such that $z^{(t)} \geqslant \upsilon$ for all $t \geqslant \mathcal{T}$. Assume that $\frac{A\upsilon^2}{m} \ll 1$. Then, we have for $\kappa \in \{1, 2\}$:

$$\prod_{\tau=0}^{\mathcal{T}-1} \left(1 + A(z^{(\tau)})^{k-1+\kappa}\right) \leqslant \left(1 + \upsilon^{k-1}\right)^{\frac{2^{\kappa+1} M A \upsilon^\kappa \log(\upsilon/z^{(0)})}{m}}.$$

Proof of Lemma K.2. Let $n \in \mathbb{N}^*$ and let T_n be the time such that $z^{(t)} \geqslant 2^n z^{(0)}$ for $t \geqslant T_n$. Starting from the recursion, we have:

$$\log(z^{(t+1)}) \geqslant \log(z^{(t)}) + \log\left(1 + m(z^{(t)})^{k-1}\right). \tag{231}$$

Since $z^{(t)}$ is a non-decreasing sequence, (231) satisfies:

$$(z^{(t+1)})^\kappa \log(z^{(t+1)}) \geqslant (z^{(t)})^\kappa \log(z^{(t)}) + (z^{(t)})^\kappa \log\left(1 + m(z^{(t)})^{k-1}\right). \tag{232}$$

We now sum (232) for $t = T_{n-1}, \dots, T_n$ and get:

$$\frac{A}{m}(z^{(T_n)})^\kappa \log(z^{(T_n)}) \geqslant \frac{A}{m}(z^{(T_{n-1})})^\kappa \log(z^{(T_{n-1})}) + \sum_{t=T_{n-1}}^{T_n-1} \frac{A(z^{(t)})^\kappa}{m} \log\left(1 + m(z^{(t)})^{k-1}\right). \tag{233}$$

Since $\frac{A(z^{(t)})^\kappa}{m} \leqslant \frac{A(2^n z^{(0)})^\kappa}{m} \leqslant \frac{A\upsilon^\kappa}{m} \ll 1$, we have $\left(1 + m(z^{(t)})^{k-1}\right)^{\frac{A(z^{(t)})^\kappa}{m}} \geqslant 1 + A(z^{(t)})^{k-1+\kappa}$. We thus lower bound (233) as:

$$\frac{A}{m}(z^{(T_n)})^\kappa \log(z^{(T_n)}) \geqslant \frac{A}{m}(z^{(T_{n-1})})^\kappa \log(z^{(T_{n-1})}) + \sum_{t=T_{n-1}}^{T_n-1} \log\left(1 + A(z^{(t)})^{k-1+\kappa}\right). \tag{234}$$

On the other hand, by using $z^{(T_n-1)} \leqslant 2^n z^{(0)}$ and $z^{(T_n)} \leqslant 2^{n+1} z^{(0)}$, we have the following upper bound.

$$\begin{aligned}
&\frac{A}{m}(z^{(T_n)})^\kappa \log(z^{(T_n)}) \\
\leqslant &\frac{A}{m}(z^{(T_n)})^\kappa \log\left(z^{(T_n-1)}\right) + \frac{MA}{m}(z^{(T_n)})^\kappa \log(1 + (z^{(T_n-1)})^{k-1}) \\
\leqslant &\frac{A \cdot (2^{n+1} z^{(0)})^\kappa}{m} \log\left(2^n z^{(0)}\right) + \frac{MA \cdot (2^{n+1} z^{(0)})^\kappa}{m} \log\left(1 + (2^n z^{(0)})^{k-1}\right).
\end{aligned} \tag{235}$$

Since $z^{(T_{n-1})} \geqslant 2^{n-1} z^{(0)}$, (235) is finally bounded as:

$$\begin{aligned}
&\frac{A}{m}\left((z^{(T_n)})^\kappa \log(z^{(T_n)}) - (z^{(T_{n-1})})^\kappa \log(z^{(T_{n-1})})\right) \\
\leqslant &\frac{\Theta(A) \cdot (2^n z^{(0)})^\kappa}{m} \log\left(2^n z^{(0)}\right) + \frac{MA \cdot (2^{n+1} z^{(0)})^\kappa}{m} \log\left(1 + (2^n z^{(0)})^{k-1}\right).
\end{aligned} \tag{236}$$

We combine (234) and (236) to obtain:

$$\begin{aligned}
\sum_{t=T_{n-1}}^{T_n-1} \log\left(1 + m(z^{(t)})^k\right) \leqslant &\frac{\Theta(A) \cdot (2^n z^{(0)})^\kappa}{m} \log\left(2^n z^{(0)}\right) \\
&+ \frac{MA \cdot (2^{n+1} z^{(0)})^\kappa}{m} \log\left(1 + (2^n z^{(0)})^{k-1}\right).
\end{aligned} \tag{237}$$

We now sum (237) and get:

$$\sum_{t=0}^{T_n-1} \log\left(1 + m(z^{(t)})^k\right) \leqslant n(2^n z^{(0)})^\kappa \left(\frac{\Theta(A)}{m} \log\left(2^n z^{(0)}\right) + \frac{2^\kappa MA}{m} \log\left(1 + (2^n z^{(0)})^{k-1}\right)\right). \tag{238}$$

We replace $2^n z^{(0)}$ by υ and n by $\log(\upsilon/z^{(0)})$ in (238) to get the aimed result. □

K.2 Probabilistic lemmas

Lemma K.3. *Let $\{\delta_r\}_{r=1}^{D-C}$ be i.i.d. random variables such that with probability q $\delta_r = \pm 1$ and zero otherwise. Then, with probability at least $1 - 1/\mathrm{poly}(d)$, we have:*

$$\sum_{\ell \neq \ell(\boldsymbol{X})} \sum_{r \in \mathcal{S}_\ell} |\delta_r| \leqslant \Theta(q(D-C)) \log(d).$$

Proof of Lemma K.3. First, note that $|\delta_r|$ is a Bernoulli random variable with parameter q. Therefore, $\sum_{\ell \neq \ell(\boldsymbol{X})} \sum_{r \in \mathcal{S}_\ell} |\delta_r|$ is a binomial random variable $\mathcal{B}(D-C, q)$. Therefore, we apply a Chernoff bound to obtain:

$$\mathbb{P}\left[\sum_{\ell \neq \ell(\boldsymbol{X})} \sum_{r \in \mathcal{S}_\ell} |\delta_r| \geqslant (1+\varepsilon)(D-C)q\right] \leqslant \exp\left(-\frac{(D-C)q\varepsilon^2}{3}\right). \tag{239}$$

Setting $\varepsilon = \sqrt{\frac{3}{(D-C)q} \log(\mathrm{poly}(d))}$ in (239) yields the desired result. □

Lemma K.4. *Let $\{\delta_r\}_{r=1}^{C}$ be i.i.d. random variables such that with probability q $\delta_r = \pm 1$ and zero otherwise. Then, with probability at least $1 - 1/\mathrm{poly}(d)$, we have:*

$$\sum_{r=1}^{C} |\delta_r| \leqslant O(1).$$

Proof of Lemma K.4. Let $k \in \mathbb{N}$ and $\Delta := \sum_{r=1}^{C} |\delta_r|$. The tail bound is bounded as:

$$\mathbb{P}[\Delta \geqslant k] = \sum_{j=k}^{C} \binom{C}{j} q^j (1-q)^{C-j} \leqslant q^k \sum_{j=k}^{C} \binom{C}{j} \leqslant q^k \sum_{j=0}^{C} \binom{C}{j} = 2^C q^k. \tag{240}$$

We want to find k such that $2^C q^k \leqslant 1/\mathrm{poly}(d)$ which implies $k \leqslant \log(d)/\log(D) \leqslant O(1)$. □

K.3 Logarithmic inequalities

Lemma K.5. *Let $a \in \mathbb{R}$ such that $C_- \leqslant a \leqslant C_+$, where $C_+, C_- > 0$. Let $p \geqslant 3$ be an odd integer. Then, the following inequality holds:*

$$\frac{0.1}{C+} \log\left(1 + \exp\left(-a^p\right)\right) \leqslant \frac{a^{p-1}}{1+\exp(a^p)} \leqslant \frac{10}{C_-} \log\left(1 + \exp\left(-a^p\right)\right).$$

Proof of Lemma K.5. We first remark that:

$$\frac{a^{p-1}}{1+\exp(a^p)} = \frac{a^p}{a(1+\exp(a^p))}. \tag{241}$$

Upper bound. We upper bound (241) by applying $a \geqslant C_-$:

$$\frac{a^{p-1}}{1+\exp(a^p)} \leqslant \frac{a^p}{C_-(1+\exp(a^p))} \tag{242}$$

We obtain the final bound by applying Lemma K.6 to (242).

Lower bound. We lower bound (241) by using $a \leqslant C_+$:

$$\frac{a^{p-1}}{1+\exp(a^p)} \geqslant \frac{a^p}{C_+(1+\exp(a^p))} \tag{243}$$

We obtain the final bound by applying Lemma K.6 to (243). □

Lemma K.6 (Connection between derivative and loss)**.** *Let $x > 0$. Then, we have:*

$$0.1 \log(1+\exp(-x)) \leqslant \mathfrak{S}(-x) \leqslant 10 \log(1+\exp(-x)) \tag{244}$$

Lemma K.7. *Let $x, y > 0$. Assume that $y \leqslant x$. Then, we have:*

$$\log(1+xy) \leqslant (1+y)\log(1+x).$$