Data Augmentation as Feature Manipulation

Ruoqi Shen [1] **Sébastien Bubeck** [2] **Suriya Gunasekar** [2]

Abstract

Data augmentation is a cornerstone of the machine learning pipeline, yet its theoretical underpinnings remain unclear. Is it merely a way to artificially augment the data set size? Or is it about encouraging the model to satisfy certain invariance? In this work we consider another angle, and we study the effect of data augmentation on the dynamic of the learning process. We find that data augmentation can alter the relative importance of various features, effectively making certain informative but hard to learn features more likely to be captured in the learning process. Importantly, we show that this effect is more pronounced for non-linear models, such as neural networks. Our main contribution is a detailed analysis of data augmentation on the learning dynamic for a two layer convolutional neural network in the recently proposed multi-view data model by Allen-Zhu & Li (2020b). We complement this analysis with further experimental evidence that data augmentation can be viewed as feature manipulation.

1. Introduction

Data augmentation is a powerful technique for inexpensively increasing the size and diversity of training data. Empirically, even minimal data augmentation can substantially increase the performance of neural networks. It is commonly argued that data augmentation is useful to impose domain specific symmetries on the model, which would be difficult to enforce directly in the architecture (Simard et al., 2000; 2003; Chapelle et al., 2001; Yaeger et al., 1996; Shorten & Khoshgoftaar, 2019). For example, semantics of a natural image is invariant under translation and scaling, so it is reasonable to augment an image data set with translated and scaled variations of its inputs. Simple augmentation with random crop up to 4 pixels can lead to gains in the range 5-10% (Ciregan et al., 2012; Krizhevsky et al., 2017). Another explanation often proposed for the role of data augmentation is merely that it increases the sample size. As an alternative to symmetry inducing or sample size increase, we consider in this work the possibility that data augmentation should in fact be viewed as a more subtle *feature manipulation* mechanism on the data.

Consider, for illustration, an image data set with the task of learning to detect whether there is a cow in the image. A simplified view would be that there are *true cow features* that generate the cow images, and we hope to learn those *true features*. At the same time, because most images of cows contain grass, it would not be surprising if a neural network would learn to detect the spurious *grass feature* for the task, and perhaps simply overfit the rare images such as desert cows that are not explained by the *grass feature* (and similarly overfit the perhaps few images with grass and no cows). Now consider a simple data augmentation technique such as Gaussian smoothing (let us assume black and white images or else use additional color space augmentations). The *grass feature*, sans color, is essentially a high frequency texture information, so we can expect the smoothing operation to make this feature significantly diminished. In this example, the *feature manipulation* that data augmentation performs is effectively to render the spurious feature harder to detect, or more precisely to make it harder to learn, which in turn boosts the *true cow features* to become the dominant features.

Continuing the illustration above, let us explore further the idea of data augmentation as *feature manipulation*. First note that the *true cow feature* need not be one "single well-defined object", but rather we may have many different true cow features. For example, *true cow features* could be different for left-facing and right-facing cows. An imbalance in the training data with respect to those different features could make the rarer features hard to learn compared to the more common features, similarly to how the spurious grass feature was occluding the true cow features. In the example above, it could happen that in most images in the training data, the cows are facing right, which in turn could mean that the neural network will learn a cow feature *with an orientation* (right-facing),

[1]University of Washington. Part of this work was done as a intern at Microsoft Research. [2]Microsoft Research. Correspondence to: Ruoqi Shen <shenr3@cs.washington.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

and then simply memorize/overfit the cows facing left. Yet another commonly used data augmentation technique such as random horizontal flip would solve this by balancing the occurrence of cow features with right-orientation and those with left-orientation, hopefully leading to a neural network dynamic that would discover *both* of those types of cow features. Note that one might be tempted to interpret this as inducing a mirror symmetry invariance in the model, but we emphasize that the effect is more subtle: the learned invariance is *only* for the relevant features, rather than being an invariance for *all* images (*e.g.,* on non-cow images one might not be invariant to the orientation).

More generally, to understand feature learning with and without data augmentation in gradient descent trained neural networks, we can think of three types of features of interest: (a) The "easy to learn and good" features, which are accurate features for the learning problem and are easy to learn in the sense that they have large relative contribution in the gradient descent updates of the network. (b) The "hard to learn and good" features, which are more nuanced to detect but are essential to fit the harder samples in the population distribution (*e.g.,* examples with rare object orientations). These are features that despite being accurate have small relative contribution in the gradient descent updates (perhaps due to lack of sufficient representation in the training data), which in turn makes them hard to learn. (c) Finally, there are the "easy to learn and bad" features, which while inaccurate, nevertheless interfere with the learning process as they have a large contribution in the gradient updates. Such features often correspond to spurious correlations or dominating noise patterns (*e.g.,* the *grass feature*) which arise due to limitations in training data size or data collection mechanisms.[1] In this paper, we study data augmentation as a technique for manipulating the "easiness" and "hardness" of features by essentially changing their relative contributions in the gradient updates for the neural network.

We believe that this view of data augmentation as a *feature manipulation* mechanism is more insightful (and closer to the truth) than the complementary and more straightforward views of "symmetry inducing" or "it's just more data". For one, data augmentation with specific symmetries do not necessarily lead to models that are respectively invariant. For example, Azulay & Weiss (2019) show that even models trained with extensive translation and scale augmentation can be sensitive to single pixel changes in translation and scaling on inputs far from the training distribution, suggesting the inductive bias from data augmentation is more subtle. Further, this view could form a basis for studying more recent data augmentation techniques like MixUp (Zhang et al., 2017), CutOut (DeVries & Taylor, 2017), and variants, which in spite of being widely successful in image tasks do not fit the conventional narrative of data augmentation.

[1] We do not mention "hard to learn and inaccurate" features as they are conceptually irrelevant for the training dynamics or accuracy of the model.

Contributions Given the diversity of data augmentation techniques (*e.g.,* see Shorten & Khoshgoftaar (2019); Feng et al. (2021) for a survey), it is a formidable challenge to understand and analyze the corresponding feature manipulation for each case, and this task is beyond the scope of the present paper. Our more modest objective is to start this program by studying a simple mathematical model where data augmentation can be provably shown to perform feature manipulation along the lines described in the illustration above. Specifically, we consider a variant of the multi-view data setting introduced in the pioneering work of Allen-Zhu & Li (2020b) on ensemble learning. In our data model, each data point is viewed as a set of patches, with each patch being represented by a high-dimensional vector in $\mathbb{R}^d$. Moreover there is a set of K "true/good" features $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K \in \mathbb{R}^d$. For any data point, each patch is then some combination of noise and features. Specifically at least one patch contains a "good" feature whose orientation indicates the label, *i.e.,* for some $k \in [K]$ this patch is $y\mathbf{v}_k$ where $y \in \{-1, 1\}$ is the binary class label to be predicted. In this case we say that the data point is of the k^{th} type. The other patches contain different forms of noise. If the training data contains sufficiently many type-k data points, then the corresponding feature $\mathbf{v}_k$ is "easy to learn and good", while the features corresponding to rare types are "hard to learn and good". To model the "easy to learn and bad" features we assume that one patch per datapoint receives a large (Gaussian) noise, which we call the dominant noise. See Section 2 for exact details of the model. Given such training data we show the following for a two layer patch-wise convolutional network (see (3)) trained using gradient descent (there is a number of caveats, see below for a list):

1. When one or more features are sufficiently rare, the network will only learn the frequent "easy to learn and good" features, and will overfit the remaining data using the "easy to learn and bad" noise component.

2. On the other hand, with any data augmentation technique that can permute or balance the features, the network will learn all K features, and thus achieve better test loss (and, importantly learn a better representation of this data[2]). We show that this happens because the representation of the "hard to learn and good" features in the gradient updates will be boosted,

[2] As a consequence of learning all the K features, the learned model will not only be more accurate on the data distribution of training samples, but will also be robust to distribution shifts that alter the proportion of data of the K feature types.

and simultaneously the relative contribution of the dominant noise or the "easy to learn and bad" features will be diminished.

3. We show that this phenomenon is more pronounced for gradient descent dynamics in non-linear models in the following sense: we prove that even at high signal-to-noise ratio (SNR) the non-linear models might memorize through the noise components, while gradient descent on linear models overfit to noise only at much lower SNR. This shows that data augmentation is useful in a wider range of cases for non-linear models than for linear models.

   Moreover, our non-linear model can learn the distribution even in the presence of feature noise (in the form of $-\alpha y \mathbf{v}_{k'}$ for some small $\alpha > 0$, which points to wrong class). On the other hand, a linear model cannot have low test error with such feature noise, thus showing a further separation between linear and non-linear models.

Some of the caveats to our theoretical results include the following points (none seem essential, but for some of them going beyond would require significant technical work):

- Neural network architecture: we study two layer neural network with a special activation function (the latter can be viewed as a smoothed ReLU with fixed bias). We also assume poly-logarithmic (in d) width.
- Training: we study gradient descent rather than stochastic gradient descent, and furthermore we assume a specific training time (the same one with and without data augmentation).
- Data model: the distribution can be generalized in many ways, including having data points with mixed types (*e.g.,* "multi-view" as in (Allen-Zhu & Li, 2020b)), heterogeneous noise components, or even correlated noise components (see below for more on this). We also assume a very high dimensional regime $d \gg n^2$ (where n is the training set size), although we believe our results should hold for $d \gg n$.

Even though our theoretical results are in a limited setting, the feature manipulation effect of data augmentation is conceptually broader. We complement our analysis with experiments on CIFAR-10 and synthetic datasets, where we study data augmentation in more generality. We circle back to our motivating problem with spurious features (*àla the cow grass features story*) in a classification task. Our experiments show that simply shifting the spurious feature position randomly up to 2 pixels in each epoch, can significantly improve the test performance by making the spurious feature hard to learn. This happens even when we do not change any non-spurious pixels/features (and hence control learning additional image priors). We further formulate experiments to evaluate the value of a single data augmented image compared to an fully independent sample, and see that on CIFAR10 dataset that once 50% independent samples are available, a data augmented sample is almost as effective as an independent sample for the learning task. Finally, we show on synthetic dataset that the problem arising from imbalance in views (as studied in our main result) also holds for deeper convolutional architectures, even when the views are merely translations of each other.

Related Work Starting with (Bishop, 1995) there is a long line of work casting data augmentation as an effective regularization technique, see (Dao et al., 2019; Rajput et al., 2019; Wu et al., 2020; Yang et al., 2022) for recent developments in that direction. Other theoretical analyses have studied and quantified the gains of data augmentation from an invariance perspective (Chen et al., 2020; Mei et al., 2021). The viewpoint we take here, based on studying directly the effect of augmentation on the learning dynamic, is strongly influenced by the work of Zeyuan Allen-Zhu and Yuanzhi Li in the last few years. For example in Allen-Zhu & Li (2020a) they develop this perspective for *adversarial training* (which in some ways can be thought as a form of data augmentation, where each data point is augmented to its adversarial version). There they show that adversarial training leads to a certain form of *feature purification*, which in essence means that the filters learned by a convolutional neural network become closer to some "ground truth" features. In (Allen-Zhu & Li, 2020b) they introduce the multi-view model that we study here, and they used it to study (among other things) ensemble learning. In a nutshell, in their version of the model each data point has several views that can be used for classification, and the idea is that each model might learn only one of those views, hence there is benefit to ensembling in that it will allow to uncover all the features, just like here we suggest that data augmentation is a way to uncover all the features. Other notable works which share the philosophy of studying the dynamic of learning (although focused on linear models) include (Hanin & Sun, 2021) which investigates the impact of data augmentation on optimization, and (Wu et al., 2020) which considers the overparametrized setting and show that data augmentation can improve generalization in this case.

Notation We use tilde notation $\widetilde{O}, \widetilde{\Theta}, \widetilde{\Omega}$ to hide $\log$ factors in standard asymptotic notation. For an integer K, $[K] = \{1, 2, \ldots, K\}$. We interchangeably use $\mathbf{a} \cdot \mathbf{b}$, $\langle \mathbf{a}, \mathbf{b} \rangle$, or $\mathbf{a}^\top \mathbf{b}$ for standard inner product between two vectors.

2. A mathematical model for understanding feature manipulation

Our data model defined below is a variation of the multi-view data distribution in (Allen-Zhu & Li, 2020b) for a binary classification task. We represent the inputs $\mathbf{x}$ as a collection of P non-overlapping patches $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_P) \in \mathbb{R}^{d \times P}$, where each patch is a d dimensional vector. The task is associated with K unknown "good" features denoted as $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K \in \mathbb{R}^d$, such that for labels $y \in \{-1, 1\}$, their orientation as $\{y\mathbf{v}_k\}_{k\in[K]}$ constitutes the K views or sub-types of the class y.[3] Each input $\mathbf{x}_p$ patches either contain one of the "good" feature $\{y\mathbf{v}_k\}$ or a "bad" feature in the form of random and/or feature noise. Formally, our distribution is defined below.

Definition 1. *$\mathcal{D}$ is parametrized by $\big(\boldsymbol{\rho}, \sigma_\xi, \sigma_\zeta, \alpha\big)$, where $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_K)$ is a discrete distribution over the features $\{\mathbf{v}_k\}_{k\in[K]}$, and σ_ξ,σ_ζ, and α are noise parameters. Without loss of generality, let $\rho_1 \geq \ldots \geq \rho_K$. A sample $(\mathbf{x}, y) \sim \mathcal{D}$ is generated as follows:*

(a) Sample $y \in \{1, -1\}$ uniformly.

(b) Given y, the input $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_P) \in \mathbb{R}^{d\times P}$ is sampled as below:

Feature patch: Choose the main feature patch $p^* \in [P]$ arbitrarily and set $\mathbf{x}_{p^*} = y\mathbf{v}_{k^*}$, where $k^* \sim \boldsymbol{\rho}$..
Dominant noise: Choose a dominant noise patch $p^\xi \neq p^*$ and generate $\mathbf{x}_{p^\xi} = \boldsymbol{\xi}$, where $\boldsymbol{\xi} \overset{i.i.d}{\sim} \mathcal{N}\big(0, \frac{\sigma_\xi^2}{d} I_d\big)$.
Background: For the remaining background patches[4] $p \in [P] \setminus \{p^*, p^\xi\}$, select $0 \leq \alpha_p \leq \alpha$ and set $\mathbf{x}_p = -\alpha_p y \mathbf{v}_{k_p} + \boldsymbol{\zeta}_p$, where $k_p \sim \boldsymbol{\rho}, \boldsymbol{\zeta}_p \sim \mathcal{N}(0, \sigma_\zeta^2 I_d)$.

Assumption 1. *We assume the features $\{\mathbf{v}_k\}_{k\in[K]}$ are orthonormal, i.e., $\forall_{k,k'\in[K]},\ \mathbf{v}_k \cdot \mathbf{v}_{k'} = \mathbf{1}_{k=k'}$.*

The training dataset consists of n i.i.d., samples from $\mathcal{D}$, $\mathcal{D}_{\text{train}} = \{(\mathbf{x}^{(i)}, y^{(i)}) : i \in [n]\} \sim \mathcal{D}^{\otimes n}$. We are interested in the high dimensional regime where $n \ll d$. n, P and K can grow with d. Note that, in Definition 1 k^*, p^*, p^ξ, $\boldsymbol{\xi}$, and $(\alpha_p, k_p, \boldsymbol{\zeta}_p)_{p\notin\{p^*,p^\xi\}}$ all depend on $\mathbf{x}$, but we have dropped this dependence in the notation to avoid clutter. In our analysis, for $i = 1, 2, \ldots, n$, we use k_i^*, p_i^*, p_i^ξ, $\boldsymbol{\xi}^{(i)}$, and $(\alpha_{p,i}, k_{p,i}, \boldsymbol{\zeta}_{p,i})_{p\notin\{p_i^*,p^\xi\}_i}$ to denote the corresponding quantities for the sample $(\mathbf{x}^{(i)}, y^{(i)})$ in the training dataset.

Data augmentation Let $\mathcal{D}_{\text{train}}^{(\text{aug})}$ denote the augmented dataset obtained by transforming the i.i.d. training dataset $\mathcal{D}_{\text{train}}$. Our model for data augmentation is such that $\mathcal{D}_{\text{train}}^{(\text{aug})}$ has equal number of samples with main feature $y\mathbf{v}_k$ for each $k \in [K]$. Concretely, consider linear transformations $\mathcal{T}_1, \ldots \mathcal{T}_{K-1}$, such that for all k, $\mathcal{T}_k : \mathbb{R}^d \to \mathbb{R}^d$ and satisfies

$$\forall\, k' \in [K],\ \mathcal{T}_k(\mathbf{v}_{k'}) = \mathbf{v}_{((k'+k-1) \bmod K)+1)}. \quad (1)$$

Such transformations are well defined for $K \leq d$, and in essence permute the feature vectors $\mathbf{v}_k$ on patches with true feature or feature noise. At the same time, the Gaussian noise patches before and after transformation are no longer i.i.d. We slightly abuse notation and define $\mathcal{T}_k(\mathbf{x})$ on $\mathbf{x} \in \mathbb{R}^{d\times P}$ as $\mathcal{T}_k(\mathbf{x}) = (\mathcal{T}_k(\mathbf{x}_1), \mathcal{T}_k(\mathbf{x}_2), \ldots, \mathcal{T}_k(\mathbf{x}_p)) \in \mathbb{R}^{d\times P}$, as well as $\mathcal{T}_k(\mathcal{D}_{\text{train}})$ on the training dataset as $\mathcal{T}_k(\mathcal{D}_{\text{train}}) = \{(\mathcal{T}_k(\mathbf{x}^{(i)}), y^{(i)}) : i \in [n]\}$.

Our augmented dataset $\mathcal{D}_{\text{train}}^{(\text{aug})}$ consists $\mathcal{D}_{\text{train}}$ along with the $K-1$ transformations of of $\mathcal{D}_{\text{train}}$ as defined below:

$$\mathcal{D}_{\text{train}}^{(\text{aug})} = \mathcal{D}_{\text{train}} \cup \mathcal{T}_1(\mathcal{D}_{\text{train}}) \ldots \cup \mathcal{T}_{K-1}(\mathcal{D}_{\text{train}}). \quad (2)$$

Note that in $\mathcal{D}_{\text{train}}^{(\text{aug})}$ all the views are equally represented, *i.e.,* for each $k \in [K]$, we will have exactly n samples from the feature $y\mathbf{v}_k$, and further $\mathcal{D}_{\text{train}}^{(\text{aug})}$ has more samples compared to $\mathcal{D}_{\text{train}}$ with $|\mathcal{D}_{\text{train}}^{(\text{aug})}| = nK$, but they are no longer i.i.d.

Since the features $\{\mathbf{v}_k\}_k$ are orthonormal (Assumption 1) and all the non-feature noise are spherically symmetric, without loss of generality, we can assume that $\{\mathbf{v}_k\}_{k\in[K]}$ are simply the first K standard basis vectors in $\mathbb{R}^d$, *i.e.,* $\mathbf{v}_k = \mathbf{e}_k$. In this case, we can choose $\mathcal{T}_k$ for $k \in [K-1]$ as a permutation of coordinates satisfying (1) on the first K coordinate. If we further assume that the the permutations $\mathcal{T}_k$ do not have any fixed points, *i.e.,* $\forall\, i \in [d]$, $\mathcal{T}_k(\mathbf{z})[i] \neq \mathbf{z}[i]$, then at initialization and updates of gradient descent, the augmented samples in $\mathcal{D}_{\text{train}}^{(\text{aug})}$ satisfy the same properties as i.i.d. samples in $\mathcal{D}_{\text{train}}$ (upto constants and log factors). In this rest of the proof, we thus assume that $\mathcal{T}_k$ are permutations of coordinates without any fixed points in the orthogonal basis extended from $\{\mathbf{v}_k\}_k$, and satisfies (1).

Role of different noise components Our main result shows that when the dominant noise parameter σ_ξ is sufficiently large, a neural network can overfit to this noise rather than learn all the views. However, with the right data augmentation, we can show that all the views can be accurately learned using a non-linear network. Furthermore, in the presence of feature noise $\{-\alpha_p y\mathbf{v}_{k_p}\}$ (pointing to wrong class), linear models are unable to fit our data distribution, thus establishing a gap from linear models.

[3] For M-class classification, our analysis can be adapted by using separate set of features $\{\mathbf{v}_{k,m}\}_k$ for each class $m \in [M]$, rather than $\{\pm\mathbf{v}_k\}_k$. For $M = 2$, under our learning algorithm, using $(\mathbf{v}_{k,-1}, \mathbf{v}_{k,1})$ as features for $y = -1, 1$ is equivalent to using $-\mathbf{v}_k, \mathbf{v}_k$ with $\mathbf{v}_k = \mathbf{v}_{k,1} - \mathbf{v}_{k,-1}$.

[4] In our definition, the dominant noise $\boldsymbol{\xi}$ and the main feature $\mathbf{v}_{k^*}$ appear in exactly one patch. But our results also hold (by virtue of parameter sharing in (3)) when for any disjoint non-empty subsets $\mathcal{P}_f, \mathcal{P}_n \subset [P]$, we set $\forall\, p \in \mathcal{P}_f,\ \mathbf{x}_p = y\mathbf{v}_{k^*}$ and $\forall\, p \in \mathcal{P}_n,\ \mathbf{x}_p = \boldsymbol{\xi}_p \underset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma_\xi^2 I_d/d)$.

We choose the noise parameters $\sigma_\xi, \sigma_\zeta, \alpha$ such that the *dominant noise* $\boldsymbol{\xi}$ and the true features $\{y\mathbf{v}_{k^*}\}$ have the main contribution to the learning dynamic compared to the feature noise (*i.e.,* $-\alpha_p y\mathbf{v}_{k_p}$) or the minor noise (*i.e.,* $\boldsymbol{\zeta}_p$). Thus, our results do not necessarily require noise in the background patches beyond establishing gap with linear models. Since the minor noise σ_ζ does not provide any additional insight, we assume $\sigma_\zeta = 0$. Our analysis can handle small σ_ζ with more tedious bookkeeping.

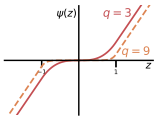
2.1. Learning algorithm

We use the following patch-wise convolutional network architecture with C channels: let $\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, \dots \mathbf{w}_C\} \in \mathbb{R}^{d\times C}$ denote the learnable parameters of the model,

$$F(\mathbf{w}, \mathbf{x}) = \sum_{c\in[C]} \sum_{p\in[P]} \psi(\mathbf{w}_c \cdot \mathbf{x}_p)\,, \tag{3}$$

where ψ is a non-linear activation function defined below:

$$\psi(z) = \begin{cases} \text{sign}(z)\cdot\frac{1}{q}|z|^q & \text{if } |z| \le 1 \\ z - \frac{q-1}{q} & \text{if } z \ge 1 \\ z + \frac{q-1}{q} & \text{if } z \le 1 \end{cases}$$



Our activation is a smoothed version of symmetrized ReLU with a fixed bias $\phi(z) = \text{ReLU}(z+1) - \text{ReLU}(-z-1)$. In fact, as $q \to \infty$, $\psi \to \phi$. Note that since we do not train the second layer weights, we choose an odd-function as activation to ensure that the outputs can be negative.

Consider the following logistic loss over the training dataset $\mathcal{D}_{\text{train}} = \{(\mathbf{x}^{(i)}, y^{(i)}), i \in [n]\}$: $L(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \ell(y^{(i)}F(\mathbf{w}, \mathbf{x}^{(i)}))$, where $\ell(z) = \log(1+\exp(-z))$.We learn the model using gradient descent on the above loss with step size η, *i.e.,* for $c \in [C]$, the weights $\mathbf{w}_c$ at time step t are given by $\mathbf{w}_c(t) = \mathbf{w}_c(t-1) - \frac{\eta}{n}\sum_{i=1}^n y^{(i)}\ell'(y^{(i)}F(\mathbf{w}(t), \mathbf{x}^{(i)}))\nabla F(\mathbf{w}(t), \mathbf{x}^{(i)})$.

The following lemma summarizes the conditions at Gaussian initialization $\mathbf{w}(0) = \{\mathbf{w}_c(0) \sim \mathcal{N}(0, \sigma_0^2 I_d) : c \in [C]\}$.

Lemma 1. *[$\mathcal{G}_{\text{init}}$-conditions] Consider n i.i.d. samples $\mathcal{D}_{train} = \{(\mathbf{x}^{(i)}, y^{(i)}) : i \in [n]\}$ from the distribution in Definition 1. Let the parameters $\mathbf{w}$ of the network in (3) be initialized as $\mathbf{w}_c(0) \sim \mathcal{N}(0, \sigma_0^2 I_d)\ \forall\, c \in [C]$. If the number of channels is $C = \Omega(\log d)$, then with probability greater than $1 - O(\frac{n^2KC}{\text{poly}(d)})$, the following conditions hold :*

1. *Feature-vs-parameter:* $\forall\, k \in [K]$, $\max_{c\in[C]} \mathbf{w}_c(0)\cdot\mathbf{v}_k \ge \Omega(\sigma_0)$, *and* $\max_{c\in[C]} |\mathbf{w}_c(0)\cdot\mathbf{v}_k| \le \widetilde{O}(\sigma_0)$.
2. *Noise-vs-parameter:* $\forall\, i \in [n]$, $\max_{c\in[C]} \mathbf{w}_c(0)\cdot y^{(i)}\boldsymbol{\xi}^{(i)} \ge \widetilde{\Omega}(\sigma_0\sigma_\xi)$, *and* $\max_{c\in[C]} |\mathbf{w}_c(0)\cdot\boldsymbol{\xi}^{(i)}| \le \widetilde{O}(\sigma_0\sigma_\xi)$.
3. *Noise-vs-noise:* $\forall\, i \in [n]$, $\boldsymbol{\xi}^{(i)}\cdot\boldsymbol{\xi}^{(i)} = \Theta(\sigma_\xi^2)$ *and* $\forall\, i, j \in [n], i \ne j$, $|\boldsymbol{\xi}^{(i)}\cdot\boldsymbol{\xi}^{(j)}| \le \widetilde{O}(\sigma_\xi^2/\sqrt{d})$.
4. *Feature-vs-noise:* $\forall\, i \in [n], k \in [K]$, $|\boldsymbol{\xi}^{(i)}\cdot\mathbf{v}_k| \le \widetilde{O}(\sigma_\xi/\sqrt{d})$.
5. *Parameter norm:* $\forall\, c \in [C]$, $\|\mathbf{w}_c(0)\| = \Theta(\sigma_0\sqrt{d})$.

The above lemma proved in the Appendix D follows from standard Gaussian concentration bounds. Further, we can show that $\mathcal{G}_{\text{init}}$ also hold for the augmented dataset $\mathcal{D}_{\text{train}}^{(\text{aug})}$ even though the samples in $\mathcal{D}_{\text{train}}^{(\text{aug})}$ are not i.i.d.

Lemma 1a. *$\mathcal{G}_{\text{init}}$ in Lemma 1 also holds for $\mathcal{D}_{train}^{(aug)}$ defined in (2) with n replaced by nK.*

2.2. Clarification on capacity in this model

We now informally discuss the size of our model class in the context of our data distribution. Consider the convolutional model (3) with $C = 1$ and say $\alpha = 0$ for sake of simplicity in the data distribution. Using $\mathbf{w}_1 = \mathbf{w}^{\texttt{gen}} = \gamma\sum_{k=1}^K \mathbf{v}_k$ for some large $\gamma > 0$ will yield excellent training and test error. This is a model that would "generalize". On the other hand for a fixed training set $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i\in[n]}$, one could also obtain almost perfect training error by using $\mathbf{w}_1 = \mathbf{w}^{\texttt{overfit}} = \gamma\sum_{i=1}^n y^{(i)}\boldsymbol{\xi}^{(i)}$, whenever σ_ξ and $d \gg n$ (noise components $\{\boldsymbol{\xi}^{(i)}\}_{i\in[n]}$ are near orthonormal). Indeed with high probability, $\forall_{i\in[n]},\ y^{(i)}f(\mathbf{w}^{\texttt{overfit}}, \mathbf{x}^{(i)}) = y^{(i)}\sum_{p\in[P]}\psi(\mathbf{w}^{\texttt{overfit}}\cdot\mathbf{x}_p^{(i)})$ is exactly

$$\psi\big(\gamma\sigma_\xi^2(1+\widetilde{O}(\sqrt{n/d}))\big) + \psi\big(\gamma\sigma_\xi O(\sqrt{n/d})\big) = \gamma\sigma_\xi^2(1+o(1))\,.$$

In other words the model with $\mathbf{w}^{\texttt{overfit}}$ will almost perfectly memorize the training set, while on the other hand it is clear that it will completely fail to generalize. This shows that the model class is large enough so that any classical measure of complexity, *e.g.,* Rademacher complexity, would fail to predict generalization (even data-dependent Rademacher complexity where the $\mathbf{x}^{(i)}$ follow our data distribution). In fact, our arguments below show that gradient descent could lead to a model of the form $\mathbf{w}^{\texttt{overfit}}$ in a Rademacher complexity setting (*i.e.,* with random label $y^{(i)}$ independent of the inputs $\mathbf{x}^{(i)}$). Thus, even restricting to models reached by gradient descent would still yield a high Rademacher complexity. This phenomenon has also been empirically observed in practical neural networks (Neyshabur et al., 2015; Zhang et al., 2021), and shown theoretically in simpler models in (Nagarajan & Kolter, 2019). Thus, we are in a case where not only do we need to leverage the fact that we are using gradient descent to prove generalization, but we also need to use the specific target function (*i.e.,* the relation between y and $\mathbf{x}$) that we are working with.

2.3. Our argument in a nutshell

At a high level we show that there is a cutoff point in the features, denote it $K_{\texttt{cut}}$, such that running gradient descent on the above architecture and data distribution will lead to a model which is essentially a mixture of parts of $\mathbf{w}^{\texttt{gen}}$ and parts of $\mathbf{w}^{\texttt{overfit}}$ described above. Roughly it will be:

$$\sum_{k \leq K_{\texttt{cut}}} \mathbf{v}_k + \sum_{i:k_i^* > K_{\texttt{cut}}} y^{(i)} \boldsymbol{\xi}^{(i)} \,. \tag{4}$$

In words, the frequent enough features will be learned, and the data points that correspond to infrequent enough features will be memorized through their noise component. Quite naturally, this cutoff point will be decreasing with the magnitude of the noise σ_ξ, *i.e.,* the bigger the noise the fewer features will be learned. While this argument also holds for gradient descent dynamics on linear models, the cutoff point $K_{\texttt{cut}}$ of linear models can be higher than that of the non-linear models, which shows that non-linear models can memorize through the noise component at a higher SNR (see Section 3.3 for the exact cutoff point).

Where data augmentation will come in is that it can effectively change the frequency of the features, and in the extreme case we consider to make them all equal,*i.e.,* all with frequency $1/K$. We then show that there exists a setting of the parameters such that frequency $1/K$ is learned at noise magnitude σ_ξ, so that with data augmentation all the features are learned.

2.4. Linear and tensor models

Before diving into the dynamics of gradient descent for our neural network architecture and data distribution, let us expand briefly on linear models. In Appendix E we study the max-ℓ_2 margin linear classifier for our data, but for sake of simplicity we consider here an even more basic predictor that is specifically tailored to our data distribution: $\bar{\boldsymbol{\theta}} := \frac{1}{n}\sum_{i=1}^{n}\sum_{p\in[P]} y^{(i)}\mathbf{x}_p^{(i)}$. Note that $\bar{\boldsymbol{\theta}}$ is a linear function on $\mathbb{R}^d$, and we naturally extend it to the domain $\mathbb{R}^{d\times P}$ of our data points (with slight overloading of notation) as $\bar{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{p\in[P]} \bar{\boldsymbol{\theta}} \cdot \mathbf{x}_p$. Compared to a gradient descent learned model, it is not clear whether this predictor is meaningful beyond our specific data distribution, and we emphasize that we study it merely as a shortest path to get quantitative estimates for the discussion in Section 2.3 (*e.g.,* for the cutoff point and for the SNR of interest). In fact the (gradient descent learnable) max margin linear classifier has even better properties than the estimator $\bar{\mathbf{w}}$, see the Appendix E for more details.

Derivation of a cutoff point. It is easy to check that with our data distribution we have $\bar{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}_S + \bar{\boldsymbol{\theta}}_N$ where $\bar{\boldsymbol{\theta}}_S = \sum_{k=1}^{K}\rho_k\mathbf{v}_k$ (say the fraction of examples of type k is exactly ρ_k) and $\bar{\boldsymbol{\theta}}_N = \frac{1}{n}\sum_{i=1}^{n} y^{(i)}\boldsymbol{\xi}^{(i)}$ (assume $\alpha = 0$ for this discussion). In particular for $\mathbf{x}$ sampled from our distribution, we have with high probability $|\bar{\boldsymbol{\theta}}_N(\mathbf{x})| \simeq \frac{\sigma_\xi^2}{\sqrt{nd}}$ and $\bar{\boldsymbol{\theta}}_S(\mathbf{x}) \simeq \rho_k y$ if $\mathbf{x}$ is of type k. This means that the predictor $\bar{\boldsymbol{\theta}}$ has successfully learned feature $\mathbf{v}_k$ iff $\rho_k > \frac{\sigma_\xi^2}{\sqrt{nd}}$. In other words for this linear model the cutoff frequency is at $\rho_{\text{cut}} = \frac{\sigma_\xi^2}{\sqrt{nd}}$. With a small leap of faith (related to the fact that after data augmentation the noise terms are no longer i.i.d., which we show to be not a in our proof of non-linear model) we can see that as long as this cutoff frequency is smaller than $\frac{1}{\sqrt{K}}$, data augmentation would enable full learning of all the views, since in that case the post-augmentation frequencies $\frac{1}{K}$ are larger than the cutoff frequency with n replaced by nK, *i.e.,* $\frac{1}{K} \gg \frac{\sigma_\xi^2}{\sqrt{nKd}} = \frac{\rho_{\text{cut}}}{\sqrt{K}}$.

Effect of simple non-linearity on SNR. The simplest type of "non-linearity" would be to consider a tensor method for this problem (note that this is nothing but a kernel method). Specifically, let $T = \frac{1}{n}\sum_{i=1}^{n}\sum_{p\in[P]} y_i \left(\mathbf{x}_p^{(i)}\right)^{\otimes q}$, be the natural empirical tensor for this problem, for some odd $q \in \mathbb{N}$, whose domain is extended from $\mathbb{R}^d$ to $\mathbb{R}^{d\times P}$ as before, i.e., $T(\mathbf{x}) = \sum_{p\in[P]} T(\mathbf{x}_p)$. Note that this function can be realized in our architecture with a pure polynomial activation function $\psi(z) = z^q$, see (Bubeck et al., 2021) for more on neural network memorization with tensors. Similarly to the linear case one can decompose the tensor into a signal and noise components:

$$T = S + N, \text{ where } S = \sum_{k=1}^{K}\rho_k\mathbf{v}_k^{\otimes q}, N = \frac{1}{n}\sum_{i=1}^{n} y_i\left(\boldsymbol{\xi}^{(i)}\right)^{\otimes q} .$$

For $\mathbf{x}$ sampled from our distribution, we have with high probability, $|N(\mathbf{x})| \simeq \frac{\sigma_\xi^{2q}}{\sqrt{nd^q}}$ and $S(\mathbf{x}) \simeq \rho_k y$ if $\mathbf{x}$ has $\mathbf{v}_k$ as its main feature. Thus here the cutoff frequency is at $\rho_{\text{cut}}^{(q)} = \frac{\sigma_\xi^{2q}}{\sqrt{nd^q}}$. In particular we see that even at high SNR, say when $\sqrt{nd} \gg \sigma_\xi^2 \gg \sqrt{d}$ (in which case $\rho_{\text{cut}}^{(1)} = o(1)$) we might have $\rho_{\text{cut}}^{(q)} = \Omega(1)$ for $q > 1$. To put it differently, the tensor methods will overfit to the noise at a different SNR from the pure linear model would, which in turns mean that there is a different range of SNR where data augmentation will be useful for non-linear models such as tensors. We will see this story repeating itself for the gradient descent on our neural network architecture.

Quantitative comparison with the neural network results. We note that the thresholds derived here are *better* than those we obtain via our neural network analysis (note also that the tensor method can handle $\alpha > 0$ similarly to what our non-linearity allows). However we emphasize again that,

on the contrary to gradient descent on neural networks, the predictors here are artificial and specifically tailored to the data distribution at hand. Furthermore the complexity of the tensor method scales up with q, on the contrary to the neural network dynamic.

3. Overview of gradient descent dynamics

Let us do some heuristic calculation in the simple case where $\alpha = 0$ (so that effectively there are only two relevant patches in inputs, $\mathbf{x}_{p^*} = y\mathbf{v}_{k^*}$ and $\mathbf{x}_{p^\xi} = \boldsymbol{\xi}$, respectively). Recall that $\mathbf{w}_c(0) \sim \mathcal{N}(0, \sigma_0^2 I_d)$ and $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_\xi^2 I_d/d)$. Thus, $\mathbb{E}[|\mathbf{w}_c(0)\cdot\mathbf{x}_{p^*}|^2] = \sigma_0^2$ and $\mathbb{E}[|\mathbf{w}_c(0)\cdot\mathbf{x}_{p^\xi}|^2] = \sigma_0^2\sigma_\xi^2$ for all channels c. We will initialize so that these quantities are $o(1)$, and thus $f(\mathbf{w}(0), \mathbf{x}) = o(1)$ for $(\mathbf{x}, y) \sim \mathcal{D}$. We study the gradient flow on minimizing f in this section.

3.1. When you really learn...

For f to correctly classify a datapoint $\mathbf{x}$ with feature $\mathbf{v}_k$, it is morally sufficient that $|\mathbf{w}_c \cdot \mathbf{v}_k|$ is of order 1 for some channel c. Let us look at the dynamics starting close to initialization (when $f(\mathbf{w}(0), \mathbf{x}) = o(1)$),

$$\begin{aligned}
&\frac{d}{dt}\mathbf{w}_c \cdot \mathbf{v}_k \\
&= -\frac{1}{n}\sum_{i\in[n]} y^{(i)}\,\ell'\big(y^{(i)}F(\mathbf{w}_c, \mathbf{x}^{(i)})\big)\left[\nabla_{\mathbf{w}_c}F(\mathbf{w}, \mathbf{x}^{(i)})\cdot\mathbf{v}_k\right] \\
&\overset{(a)}{=} \frac{1+o(1)}{2n}\sum_{i\in[n]}\sum_{p\in[P]}\psi'(|\mathbf{w}_c\cdot\mathbf{x}_p^{(i)}|)\,y^{(i)}\mathbf{x}_p^{(i)}\cdot\mathbf{v}_k \\
&= \frac{1+o(1)}{2n}\sum_{i\in[n]}\psi'(|\mathbf{w}_c\cdot\mathbf{v}_{k_i^*}|)\mathbf{v}_{k_i^*}\cdot\mathbf{v}_k + \\
&\underbrace{\frac{1+o(1)}{2n}\sum_{i\in[n]}\psi'(|\mathbf{w}_c\cdot\boldsymbol{\xi}^{(i)}|)y^{(i)}\boldsymbol{\xi}^{(i)}\cdot\mathbf{v}_k}_{:=\vartheta} \\
&\overset{(b)}{=} \frac{1+o(1)}{2}\,\rho_k\,\psi'(|\mathbf{w}_c\cdot\mathbf{v}_k|) + \vartheta\,, \qquad (5)
\end{aligned}$$

where in (a), we use $-\ell'(o(1)) = 1/2 + o(1)$ for logistic loss ℓ, $\psi'(z) = \psi(|z|)$ since ψ is odd, and (b) follows from $\{\mathbf{v}_k\}$ being orthogonal.

If we can ignore ϑ, resulting dynamic reduces to an ODE of the form $g'(t) = \rho_k\psi'(g(t))$ (ignoring constants) with $g(0) \approx \sigma_0 = o(1)$. As long as $g(t) = \mathbf{w}_c(t)\cdot\mathbf{v}_k$ is smaller than 1 this can be rewritten as $g'(t) = \rho_k g(t)^{q-1}$ (because of the form of ψ we chose), or equivalently $(g(t)^{2-q})' = -\rho_k$ up to constants. In particular, we see that after time $t = g(0)^{2-q}/\rho_k$, we will have $g(t) = \Theta(1)$. This suggests that by time of order $1/(\sigma_0^{q-2}\rho_k)$ at least one channel should have learned $\mathbf{v}_k$[5].

[5]We assume $q \geq 3$. For the case $q = 1$ or $q = 2$, the time needed is $1/(\sigma_0^{q-1}\rho_k)$.

When can we indeed ignore (morally) the noise term ϑ*?* At initialization this term is of order $\frac{\sigma_0^{q-1}\sigma_\xi^q}{\sqrt{nd}}$. On the other hand the "main" term $\mathbf{w}_c\cdot\mathbf{v}_k$ in (5) is of order $\rho_k\sigma_0^{q-1}$. Thus we see that we need $\frac{\sigma_\xi^q}{\sqrt{nd}} \ll \rho_k$. In fact we will need a slightly more stringent condition, because the cancellation in ϑ leading to a scaling of $1/\sqrt{n}$ becomes more complicated to analyze after initialization due to the dependencies getting introduced. So we will use the more brutal bound $|\vartheta| \lesssim \frac{\sigma_0^{q-1}\sigma_\xi^q}{\sqrt{d}}$ which in turn means we need $\frac{\sigma_\xi^q}{\sqrt{d}} \ll \rho_k$.

Summarizing the above, we expect that if $\sigma_\xi^q/\sqrt{d} \ll \rho_k$, then by time $1/(\sigma_0^{q-2}\rho_k)$ we will have one channel that has learned the feature $\mathbf{v}_k$.

3.2. ... and when you overfit ...

Another sufficient condition to correctly classify a datapoint $(\mathbf{x}^{(j)}, y^{(j)})$ would be to overfit to its dominant noise part $\boldsymbol{\xi}^{(j)}$, *i.e.,* $|\mathbf{w}_c\cdot\boldsymbol{\xi}^{(j)}|$ is of order 1 for some channel c. Here we have at initialization:

$$\begin{aligned}
\frac{d}{dt}\mathbf{w}_c\cdot\boldsymbol{\xi}^{(j)} &= \frac{1+o(1)}{2n}\sum_{i\in[n]}\sum_{p\in[P]}\psi'(|\mathbf{w}_c\cdot\mathbf{x}_p^{(i)}|)y^{(i)}\mathbf{x}_p^{(i)}\cdot\boldsymbol{\xi}^{(j)} \\
&= \frac{1+o(1)}{2n}\Bigg(y^{(j)}\psi'(|\mathbf{w}_c\cdot\boldsymbol{\xi}^{(j)}|)\|\boldsymbol{\xi}^{(j)}\|^2 \\
&+ \psi'(|\mathbf{w}_c\cdot\mathbf{v}_{k_j^*}|)\mathbf{v}_{k_j^*}\cdot\boldsymbol{\xi}^{(j)} \\
&+ \sum_{i\neq j, p\in[P]}\psi'(|\mathbf{w}_c\cdot\mathbf{x}_p^{(i)}|)\,y^{(i)}\mathbf{x}_p^{(i)}\cdot\boldsymbol{\xi}^{(j)}\Bigg) \\
&= \frac{(1+o(1))\sigma_\xi^2}{2n}y^{(j)}\psi'(|\mathbf{w}_c\cdot\boldsymbol{\xi}^{(j)}|) + \Gamma \qquad (6)
\end{aligned}$$

where Γ is the last two term from the penultimate step.

Assuming Γ *can be ignored,* we can mimic the reasoning above (for $\mathbf{w}_c\cdot\mathbf{v}_k$) with $h(t) = y^{(j)}\mathbf{w}_c\cdot\boldsymbol{\xi}^{(j)}$ and $h(0) = O(\sigma_0\sigma_\xi)$. We thus expect to correctly classify a datapoint by *overfitting to its noise* after time $O(n/(\sigma_0^{q-2}\sigma_\xi^q))$.

When can we ignore the noise term Γ*?* The order of Γ is $\sigma_\xi^{q+1}\sigma_0^{q-1}/\sqrt{d}$ (at initialization it is in fact this times $1/\sqrt{n}$ but we ignore this improvement due to the dependencies arising through learning). On the other hand the main term in (6) is of order $\sigma_\xi^{q+1}\sigma_0^{q-1}/n$ at initialization, so we obtain the condition $\sqrt{d} \gg n$ (which could possibly be improved to $d \gg n$ if cancellation remained correct throughout learning).

Summarizing again, if $d \gg n^2$, by time in the order of $n/(\sigma_0^{q-2}\sigma_\xi^q)$, we can expect the data points that were not fit before this time to be overfit using noise parameters.

3.3. ... and in what order

Let us assume $d \gg n^2$ and $\frac{\sigma_\xi^q}{\sqrt{d}} \ll \rho_k$. Then the above discussion reveals that if $n/(\sigma_0^{q-2}\sigma_\xi^q) \ll 1/(\sigma_0\rho_k) \Leftrightarrow \rho_k \ll \sigma_\xi^q/n$, we will not be able to learn $\mathbf{v}_k$ because we will overfit before learning (In fact, in this case, we do not need the condition $\frac{\sigma_\xi^q}{\sqrt{d}} \ll \rho_k$). This essentially gives rise to a channel filter (or a combination thereof) of the form (4), with the cutoff point $K_{out} = \{k : \rho_k \ll \sigma_\xi^q/n\}$ being now specified.

Data augmentation can fix the order by effectively permuting the features. After data augmentation, we get the proportion of any feature to be $1/K$ and the training set size to be nK. Note that our data augmentation only permutes the coordinates so that the inner product between $\boldsymbol{\xi}$ and $\mathcal{T}_k(\boldsymbol{\xi})$ should be at the same order as two independent noise. The learning process only depend on the inner product between the samples so our previous analysis still holds. Then, after data augmentation, for every view $k \in [K]$, we have $\rho_k^{(\text{aug})} = 1/K$. Then, as long as $\sigma_\xi^q/n = o(1)$, we have $\rho_k^{(\text{aug})} \gg \sigma_\xi^q/(nK)$ and are able to learn $\mathbf{v}_k$ before overfitting.

3.4. What about spurious features?

In addition to overfitting noise, the model can also overfit spurious features. The spurious features can be viewed as noise vectors that appear in more than one sample. We do not prove this case formally in our main theorems for simplicity, but we will give the proof intuition here. Let $\mathbf{u} \in \mathbb{R}^d$, $\|\mathbf{u}\| = 1$, be some spurious feature. Now assume that in addition to the dominant noise patch and the feature patch, $\mathbf{u}$ appears in $1 > \rho_{\mathbf{u}}^{(-1)} > 0$ fraction of the datapoints with label $y = 1$ and $\rho_{\mathbf{u}}^{(-1)} < \rho_{\mathbf{u}}^{(1)}$ fraction of the datapoints with label $y = -1$. We assume $\mathbf{u}$ is orthogonal to the main features $\mathbf{v}_1, ..., \mathbf{v}_K$. Let $\mathcal{I}_{\mathbf{u}}$ be the set of samples with $\mathbf{u}$. We have at initialization:

$$\begin{aligned}
&\frac{d}{dt}\mathbf{w}_c \cdot \mathbf{u} \\
&= \frac{1+o(1)}{2n}\sum_{i\in[n]}\sum_{p\in[P]}\psi'(|\mathbf{w}_c \cdot \mathbf{x}_p^{(i)}|)y^{(i)}\mathbf{x}_p^{(i)} \cdot \mathbf{u} \\
&= \frac{1+o(1)}{2n}\sum_{i\in\mathcal{I}_{\mathbf{u}}} y^{(i)}\psi'(|\mathbf{w}_c \cdot \mathbf{u}|)\|\mathbf{u}\|^2 \\
&+ \underbrace{\frac{1+o(1)}{2n}\sum_{i\in[n]}\psi'(|\mathbf{w}_c \cdot \boldsymbol{\xi}^{(i)}|)y^{(i)}\boldsymbol{\xi}^{(i)} \cdot \mathbf{u}}_{:=\Upsilon} \\
&= \frac{1+o(1)}{2n}(\rho_{\mathbf{u}}^{(1)} - \rho_{\mathbf{u}}^{(-1)})\psi'(|\mathbf{w}_c \cdot \mathbf{u}|)\|\mathbf{u}\|^2 + \Upsilon. \quad (7)
\end{aligned}$$

Assuming Υ can be ignored, we can mimic the reasoning for $\mathbf{w}_c \cdot \mathbf{v}_k$ with $h(t) = \mathbf{w}_c \cdot \mathbf{u}$ and $h(0) = O(\sigma_0)$. We thus expect to correctly classify a datapoint in class $y = 1$ with spurious feature $\mathbf{u}$ by *overfitting to* $\mathbf{u}$ after time $O(n/(\sigma_0^{q-2}(\rho_{\mathbf{u}}^{(1)} - \rho_{\mathbf{u}}^{(-1)})))$.

When can we ignore the noise term Υ? Similar to the term ϑ in (5), the order of Υ is $\sigma_\xi^q\sigma_0^{q-1}/\sqrt{nd}$. On the other hand the main term in (7) is of order $(\rho_{\mathbf{u}}^{(1)} - \rho_{\mathbf{u}}^{(-1)})\sigma_0^{q-1}$. Thus we need $\frac{\sigma_\xi^q}{\sqrt{nd}} \ll \rho_{\mathbf{u}}^{(1)} - \rho_{\mathbf{u}}^{(-1)}$.

Summarizing above, since $\mathbf{u}$ can appear in both class $y = 1$ and class $y = -1$, it should not be used as an indicator of the label y. However, when $\mathbf{u}$ appears predominantly in one class (e.g., when $\frac{\sigma_\xi^q}{\sqrt{nd}} \ll \rho_{\mathbf{u}}^{(1)} - \rho_{\mathbf{u}}^{(-1)}$), the model can overfit $\mathbf{u}$ and use $\mathbf{u}$ to classify the datapoints.

4. Main Results

We learn the model $F(\mathbf{w}, \mathbf{x})$ in (3) using gradient descent with step size η on loss $L(\mathbf{w})$ in (**??**). The weight $\mathbf{w}_c$, $c \in [C]$, at time step t is denoted as $\mathbf{w}_c(t)$. The weight $\mathbf{w}_c(t)$ for training on $\mathcal{D}_{\text{train}}^{(\text{aug})}$ is obtained similarly, with the samples replaced by $\mathcal{D}_{\text{train}}^{(\text{aug})} = \{(\mathbf{x}^{(i)}, y^{(i)}), i \in [Kn]\}$. In addition to the assumptions we have discussed in Section 3, we make some additional assumptions for controlling the omitted quantities arising through training and testing.

Assumption 2. *We assume the following holds. For some constant $q \geq 3$,*

1. *The first view is dominant, $1 \geq \rho_1 \geq \Omega(1)$. The other views $k \in [K]\backslash\{1\}$ are minor views, $n\rho_k \leq o\left(\sigma_\xi^q\right)$.*
2. *The standard deviation of the dominant noise satisfies $\omega(1) \leq \sigma_\xi^q \leq o(n)$.*
3. *The standard deviation of the weights at initialization is bounded, $\sigma_0 \leq o(1/\sigma_\xi)$.*
4. *The number of samples and views are bounded, $nK \leq o\left(\sigma_0^{q-1}\sigma_\xi^{q-1}d^{1/2}\right)$.*
5. *The feature noise satisfies, for $T = \widetilde{\Theta}\left(\max\left\{n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}, K\eta^{-1}\sigma_0^{-q+2}\right\}\right)$,*

$$\omega(P^{-1}) \leq \alpha \leq o\left(\eta^{-1}T^{-1}P^{-\frac{1}{q}}\sigma_\xi \min\{d^{-\frac{1}{2}}, \sigma_0\}\right).$$

Condition 1-3 in Assumption 2 have been explained in Section 3. $\sigma_0 \leq o(1)$ and $\sigma_0\sigma_\xi \leq o(1)$ guarantee that at initialization, the main features and the dominant noise have $o(1)$ correlation with the weight. We choose $\sigma_\xi \geq \omega(1)$ so that without properly learning the main feature, the inner product between random initialized weights and the dominant patch can dominate the model output. Condition 4 is a more stringent version of the condition $n \ll d^{1/2}$

in Section 3 to control all the terms during training. In Condition 5, we assume an upper bound on the feature noise α. We assume the existence of feature noise only for establishing gap with linear models, so we did not optimize the upper bound on α. It is possible the proof can go through with milder constraints on α.

An example of a set of parameters that satisfy the above assumption is

$$q = 3, \sigma_0 = d^{-0.15}, \sigma_\xi = d^{0.1}, n = d^{0.33},$$
$$K = d^{0.06}, \rho_1 = \frac{1}{2}, \rho_2 = \rho_3 = ... = \rho_K = \frac{1}{2(K-1)},$$
$$\alpha = d^{-0.95}, P = d.$$

In Theorem 3, we show that under the above conditions, without data augmentation, gradient descent can find a classifier with perfect training accuracy without learning the minor views. On the other hand, Theorem 4 shows that with data augmentation, all k views can be learned without overfitting to noise.

Theorem 3 (Training without data augmentation)**.** *Suppose that Assumption 2 holds. Let $\overline{T}$ be the first time step such that $\mathbf{w}(\overline{T})$ can classify all $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}_{train}$ with constant margin, i.e., ,*

$$y^{(i)} F(\mathbf{w}(\overline{T}), \mathbf{x}^{(i)}) \geq \widetilde{\Omega}(1), \quad \textit{for all } (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}_{train}.$$

For hidden channel number $C = \Theta(\log d)$, and small step size η, with probability at least $1 - O(\frac{n^2 K}{\text{poly}(d)})$, $\overline{T} = \widetilde{\Theta}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$. Moreover, at time step $\overline{T}$, views $\mathbf{v}_2, \ldots, \mathbf{v}_K$ have never been learned, so that $\forall_{0 \leq t \leq \overline{T}}$,

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}}[yF(\mathbf{w}(t), \mathbf{x}) < 0] \geq \left(\frac{1}{2} - O\left(\frac{1}{\sqrt{C}}\right)\right) \sum_{k=2}^{K} \rho_k.$$

Theorem 4 (Training with data augmentation)**.** *Suppose assumption 2 holds. Let $\overline{T}_{aug}$ be the first time step such that $\mathbf{w}(\overline{T}_{aug})$ can classify all $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}_{train}^{(aug)}$ with constant margin, i.e.,*

$$y^{(i)} F(\mathbf{w}(\overline{T}_{aug}), \mathbf{x}^{(i)}) \geq \widetilde{\Omega}(1), \quad \textit{for all } (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}_{train}^{(aug)}.$$

For hidden channels number $C = \Theta(\log d)$, and small step size η, with probability at least $1 - O(\frac{n^2 K^3}{\text{poly}(d)})$, $\overline{T}_{aug} = \widetilde{\Theta}\left(K\eta^{-1}\sigma_0^{-q+2}\right)$, and at $\overline{T}_{aug}$,

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}}\left[yF(\mathbf{w}(\overline{T}_{aug}), \mathbf{x}) < 0\right] \leq \frac{nK}{\text{poly}(d)}.$$

Remark 5. In Theorem 3 and Theorem 4, we evaluate the testing accuracy at the earliest time step T when the trained neural network with weights $\mathbf{w}(T)$ can classify all samples in the training set $\mathcal{D}_{\text{train}}$ with a constant margin. Our result does not rule out the possibility that if trained longer than $\bar{T}$, the network can learn the minor views as well. However, we should expect the gradients on the training set stay small after the network can classify all sample correctly. The main reason we assume an upper bound on ηT is when training too long, the norm of the weights $\mathbf{w}$ can blow up. One possible strategy to avoid such upper bound on ηT is to add weight decay to the gradient descent algorithm in training.

Remark 6. For simplicity of the proof, we only keep track of the channel with the maximum correlation with the main feature or the noise, $\arg\max_{c\in[C]} \mathbf{w}_c(t) \cdot \mathbf{v}_k$ and $\arg\max_{c\in[C]} y\mathbf{w}_c(t) \cdot \boldsymbol{\xi}$. For the other channels, we only give a rough bound on their correlation. For this reason, we assume the number of channels is $C = \Theta(\log d)$ so that the output is dominated by the channel with the maximum correlation. To extend the result to higher number of channels, such as polynomial in d, we need to keep track of all channels and scale the output layer by $\frac{1}{C}$.

Remark 7. In our model, we show that when there exists some large dominant noise, the neural network overfits to the noise instead of learning the minor features. In practice, the model can overfit to any vector that contributes significantly to the gradient of the loss. For example, our proof can be extended to the case where there exists some spurious feature that appears in sufficiently many samples. In such case, even when the magnitude of the spurious feature is smaller than the dominant noise in our distribution, the network can still overfit it.

5. Experiment

Our theoretical results showed that data augmentation can make it harder to overfit to the noise components (the "easy to learn and bad" feature in our model) by manipulating the relative gradient contribution of noise vs true features. To simplify our analysis, we assumed independent dominant noise in each sample. We hypothesize that the feature manipulation effect of data augmentation is broader in practice. In particular, our high level argument suggests that a model can also overfit to spurious features, like the *grass feature* in our story of cows in the introduction, which have strong class dependent correlations. In Appedix A, we complement our theory using experiments.

Acknowledgements

We would like to thank Yi Zhang for valuable feedback.

References

Allen-Zhu, Z. and Li, Y. Feature purification: How adversarial training performs robust deep learning. *arXiv preprint arXiv:2005.10190*, 2020a.

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020b.

Azulay, A. and Weiss, Y. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20:1–25, 2019.

Berry, A. C. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941.

Bishop, C. M. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

Bubeck, S., Li, Y., and Nagaraj, D. A law of robustness for two-layers neural networks. *Conference on Learning Theory (COLT)*, 2021.

Chapelle, O., Weston, J., Bottou, L., and Vapnik, V. Vicinal risk minimization. *Advances in neural information processing systems*, pp. 416–422, 2001.

Chen, S., Dobriban, E., and Lee, J. H. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

Ciregan, D., Meier, U., and Schmidhuber, J. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3642–3649. IEEE, 2012.

Dao, T., Gu, A., Ratner, A., Smith, V., De Sa, C., and Ré, C. A kernel theory of modern data augmentation. In *International Conference on Machine Learning*, pp. 1528–1537. PMLR, 2019.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Elandt, R. C. The folded normal distribution: Two methods of estimating parameters from moments. *Technometrics*, 3(4):551–562, 1961.

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., and Hovy, E. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, 2021.

Hanin, B. and Sun, Y. How data augmentation affects optimization for linear regression. *Advances in Neural Information Processing Systems*, 34, 2021.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Mei, S., Misiakiewicz, T., and Montanari, A. Learning with invariances in random features and kernel models. *arXiv preprint arXiv:2102.13219*, 2021.

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/05e97c207235d63ceb1db43c60db7bbb-Paper.pdf.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR (Workshop)*, 2015.

Rajput, S., Feng, Z., Charles, Z., Loh, P.-L., and Papailiopoulos, D. Does data augmentation lead to positive margin? In *International Conference on Machine Learning*, pp. 5321–5330. PMLR, 2019.

Shorten, C. and Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

Simard, P. Y., Le Cun, Y. A., Denker, J. S., and Victorri, B. Transformation invariance in pattern recognition: Tangent distance and propagation. *International Journal of Imaging Systems and Technology*, 11(3):181–197, 2000.

Simard, P. Y., Steinkraus, D., and Platt, J. C. Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, volume 3, pp. 958–958. IEEE Computer Society, 2003.

Wu, S., Zhang, H., Valiant, G., and Ré, C. On the generalization effects of linear transformations in data augmentation. In *International Conference on Machine Learning*, pp. 10410–10420. PMLR, 2020.

Yaeger, L., Lyon, R., and Webb, B. Effective training of a neural network character classifier for word recognition. *Advances in neural information processing systems*, 9: 807–816, 1996.

Yang, S., Dong, Y., Ward, R., Dhillon, I. S., Sanghavi, S., and Lei, Q. Sample efficiency of data augmentation consistency regularization. *arXiv preprint arXiv:2202.12230*, 2022.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

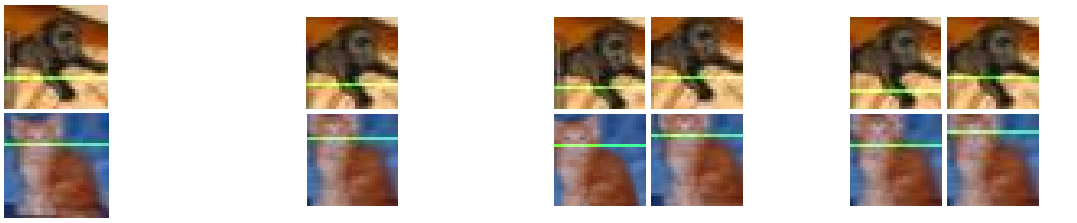
A. Experiments

Our theoretical results showed that data augmentation can make it harder to overfit to the noise components (the "easy to learn and bad" feature in our model) by manipulating the relative gradient contribution of noise vs true features. To simplify our analysis, we assumed independent dominant noise in each sample. We hypothesize that the feature manipulation effect of data augmentation is broader in practice. In particular, our high level argument suggests that a model can also overfit to spurious features, like the *grass feature* in our story of cows in the introduction, which have strong class dependent correlations. In Section A.1, we show experiments to this effect that complement our theory. We further conduct two additional experiments that support this paper's thesis. In Section A.2, we show an experiment with a modified data augmentation pipeline that demonstrates that the benefits of data augmentation cannot be fully explained by the learning of right invariance by the model. Finally, in Section A.3 we elaborate on the problem with unbalanced views, where we show that adding extra samples from one dominant view to balanced dataset can hurt the performance of the learned models.

A.1. Spurious Feature

We use images of the dog class and the cat class from CIFAR-10 dataset, which are of size 32×32 pixels and 3 channels. We generate a row of random pixels $\mathbf{u} \sim \mathcal{N}(0, \sigma^2 I_d)$, where $d = 32$ and $\sigma = 25$, which is added as a synthetic spurious feature to a class dependent position in an image. The spurious feature $\mathbf{u}$ is added to the first channel in the row r_{cat} for cat images, and in row r_{dog} for dog images. For each image $\mathbf{x}$ in the dataset, with probability $p < 1$ we introduce a spurious feature, and with probability $(1 - p)$ we leave it unperturbed. We always select $r_{\text{cat}} \in \{0, 1, \ldots, 15\}$ in the upper half of the image, and $r_{\text{dog}} \in \{16, 17, \ldots, 31\}$ in the lower half. In this way, the spurious feature position has a weak correlation to the class label. See Figure 1 for sample images with spurious features. We consider three types training sets with varying degrees of data augmentation as shown Figure 1-(b,c,d).

1. *No augmentation:* As a baseline without augmentation, we center-crop the image to size $[3, 28, 28]$.
2. *Random crop:* In each epoch, we randomly crop a $[3, 28, 28]$ from the original $[3, 32, 32]$ image—a standard technique used in practice. This would in essence disperse the position of spurious feature $\mathbf{u}$. For example, cat images with $\mathbf{u}$ in row $r_{\text{cat}} = 9$, will now contain $\mathbf{u}$ in a row uniformly chosen from $r_{\text{cat}}^{\text{aug}} \sim \mathcal{U}(\{5, 6, 7, 8, 9\})$.
3. *Randomized noise position:* Random crop, although standard, has a confounding effect that in addition to perturbing the position of $\mathbf{u}$, it might also incorporate other useful inductive biases about images. For a more direct comparison to the baseline, we also look at a special augmentation, wherein we perturb just the spurious feature row position by a uniform random number in $[-2, 2]$ in each epoch and then use a simple center crop. As in the case of random crop, this would again disperse the spurious feature from $r_{\text{cat}} = 9$ to $r_{\text{cat}}^{\text{aug}} \in \mathcal{U}(\{5, 6, 7, 8, 9\})$. But the non-spurious features/pixels remain the same as baseline.



(a) Original images (b) No augmentation (c) Random crop (d) Random noise position

Figure 1: Examples of training images in the spurious features experiment (Section A.1). For ease of visualization, we use a green line rather than random row vector $\mathbf{u}$ to indicate the spurious feature. In the original $[3, 32, 32]$ images shown in (a), the spurious feature is added to the first channel of row $r_{\text{cat}} = 9$ for the cat class (lower images), and of row $r_{\text{dog}} = 22$ for the dog class (upper images). Sub-figures (b,c,d) correspond to samples from different data augmentation methods described in the experiment.

We compare the testing accuracy of training on these three types of training set in Figure 2 for different values of r_{cat} and r_{dog}. When $(r_{\text{cat}}, r_{\text{dog}}) = (15, 16)$ (Figure 2, right), after data augmentation with either *random noise position* or *random crop*, the position of $\mathbf{u}$ in the perturbed imaged has a large overlap across classes. So it is not surprising that the test accuracy with augmentation remains about the same for almost all values of p (fraction of images with spurious features). On the other hand, for positions $(9, 22)$ and $(12, 19)$ (Figure 2, left & center)), although the two data augmentation techniques disperse the positions of spurious feature, its location in the two classes still stays separated. The cat images always have $\mathbf{u}$ in the upper half of the image while the dog images always have $\mathbf{u}$ in the lower half of the image. Interestingly, even so, the

data augmentation, specially even the simple *random feature position*, can improve the test accuracy. In this case, while augmentation does not eliminate the existence of spurious features, it still diminishes them by making the spurious features harder to be learned and overfitted. In addition to shifting the spurious features, random crop can shift other important features as well to boost the minor views, so the testing accuracy when training with random crop can be even higher than only shifting the spurious feature position.

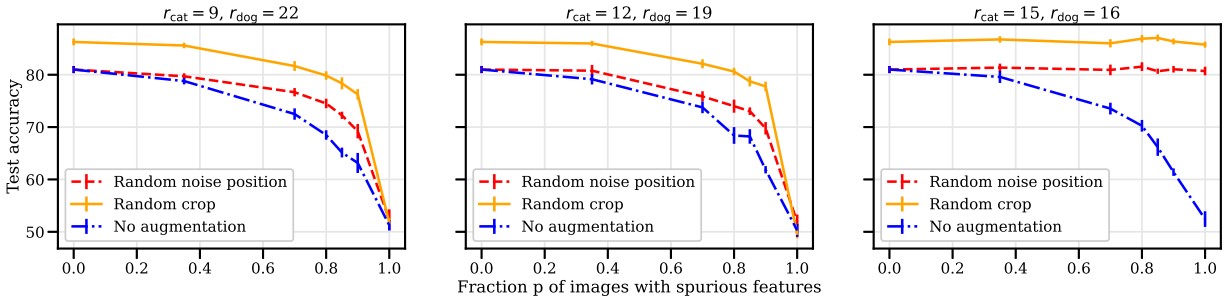


Figure 2: Comparison of different data augmentation strategies for the CIFAR-10 cat-vs-dog classification task with a synthetic spurious feature. The plots show results for different sets of positions of spurious feature $(r_{\text{cat}}, r_{\text{dog}})$ as we vary the fraction p of all the images that have the spurious feature. The plots are averaged over five runs with error bars of one standard deviation. The test datapoints are always center-cropped images of size $[3, 28, 28]$ with no spurious feature. In all configurations, we train a ResNet20 network using SGD for 120 epochs with momentum 0.9, weight decay 0.005, and learning rate starting at 0.1 and annealed to $(0.01, 0.001)$ at epochs $(40, 80)$.

A.2. Augmented samples vs. independent samples

When using data augmentation, typically a new random transformation (*e.g.,* random flip or crop at a random position of an image) is used in each epoch of training. This procedure effectively increases the training dataset size (albeit with non i.i.d correlated samples). In this experiment, we control for the number of unique samples seen by the training algorithm and ask the question: *how effective is a single data augmented sample compared to an independent sample?*

For this experiment, we work with the full CIFAR-10 dataset which has 50000 training examples for 10 classes. Given a ratio p of independent samples to total sample size, we generate a training set of size $n = 50000$ as follows: We first select pn independent samples for the task. We then cyclically generate a data augmented variant these pn independent samples until we obtain the remaining $(1 - p)n$ datapoint. For example, in the CIFAR-10 dataset with $n = 50000$, if $p = 0.6$, the training set consists of 30000 independent samples, of which 20000 have one additional augmented sample. If $p = 0.2$, the training set has 10000 independent samples and four data augmented versions of each of the 10000 independent samples. Thus, for $p = 1$, there is no augmentation, and for smaller p, there are more augmented samples, but less independent samples. The dataset thus generated is then fixed for all epochs. In this way, the number of unique samples seen by training algorithm is always $n = 50000$ for all p.

In Figure 3, we compare the accuracies of a ResNet20 model trained on such partially data augmented samples to the baseline of training with just the pn independent samples without any augmentation. Our experiment shows that even this partial data augmentation can significantly improve the testing accuracy. In this experiment, since each example has only a small number of augmented variations (*e.g.,* for $p \geq 0.5$ at most one augmented variant of the an example is seen throughout training), it is unlikely that they lead to learning any kind of task specific invariance, which is the usual motivation. However, by having the important feature appearing at a slightly different location, data augmentation can still facilitate the learning of the important features via the feature manipulation view described in our paper. Furthermore, comparing the accuracy of un-augmented full dataset with $p = 1.0$ on blue-dashed curve to that of data augmented training for $p \geq 0.5$ on the red curve, we see that a fixed data augmented image can improve the test accuracy nearly as much as an independent sample does. This shows that if we have an important feature in an image, *e.g.,* a cat ear, shifting it two pixels can help nearly as effectively as a completely new cat ear.

A.3. Unbalanced Dataset

In this experiment, we train a simple convolutional neural network on a synthetic dataset with unbalanced views. We show that when one view is much more prevalent in the dataset than the other views, having more samples of the dominant view can hurt learning. Our data consist of samples $(\mathbf{x}, y)$ from two classes $y \in \{-1, 1\}$. The input $\mathbf{x} \in \mathbb{R}^{3\times 15}$ has 3 channels,

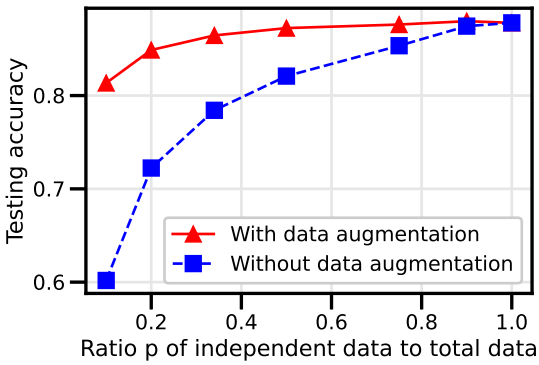


Figure 3: Augmented vs independent samples: for each p on the x-axis, the data augmented training (red-solid curve) uses $50000p$ independent images from CIFAR-10, along with $50000(1-p)$ data augmented samples. The augmented dataset is fixed across epochs. For the baseline without data augmentation (blue-dashed curve) we simply use the $50000p$ independent samples. We use the standard CIFAR-10 test dataset and the results are averaged over 3 runs. In each instance, we train a ResNet20 for 160 epochs using SGD with momentum 0.9, weight decay 0.005, and learning rate starting at 0.1 and annealed to $(0.01, 0.001)$ at epochs $(80, 120)$.

each with 15 pixels. After sampling y uniformly, we generate $\mathbf{x}$ by setting one of the 15 pixels to the main feature $[y, y, y]$. The other pixels are set to a Gaussian noise $\mathcal{N}(0, \sigma_\xi^2 I_3)$. For different choices of σ_ξ, we first construct a balanced dataset $\mathcal{D}_{\text{bal}}$ of size n_{bal} such that roughly equal number of samples that have the good feature $[y, y, y]$ present at each pixel. Our full dataset $\mathcal{D}_{\text{full}}$ with n_{full} samples consists of $\mathcal{D}_{\text{bal}}$ along with additional $n_{\text{full}} - n_{\text{bal}}$ samples with the main feature only at pixel 3. We use a balanced testing dataset.

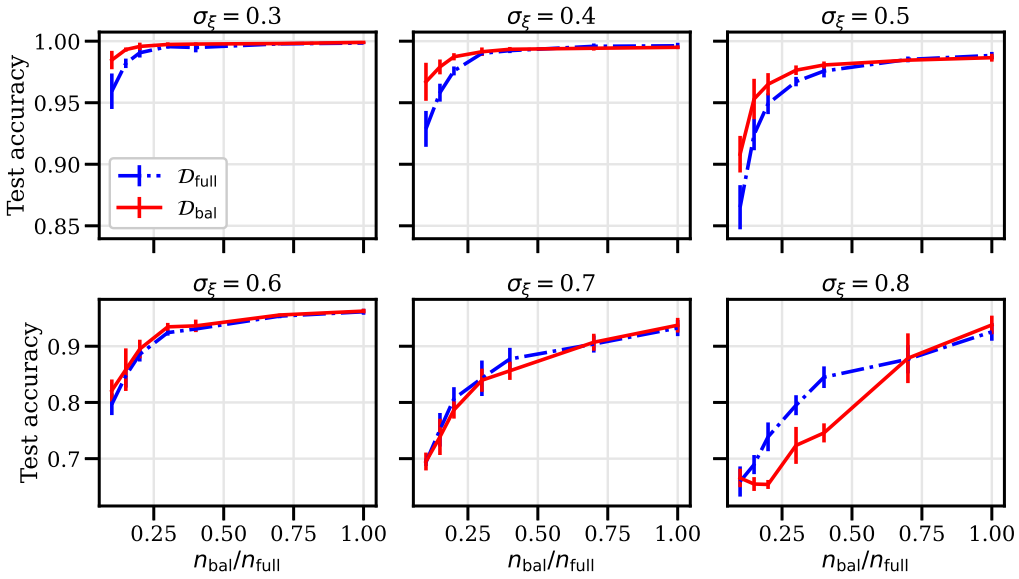


Figure 4: Comparison of training on $\mathcal{D}_{\text{bal}}$ to $\mathcal{D}_{\text{full}}$ as we vary the ratio of balanced examples $n_{\text{bal}}/n_{\text{full}}$ for different values of noise magnitude σ_ξ. We learn the data using a simple convolutional neural network with two convolutional layers with ReLU activation, a maxpool layer and a linear layer. The two convolutional layers and the max pool layer have kernel size 4, and strides 2,1 and 2, respectively. The models are trained for 200 epochs using SGD with momentum 0.9, weight decay 0.05, and learning rate starting at 0.1 and annealed to 0.01 at epoch 100. For all training sets, the training accuracy at the end of training is at least 0.99.

In Figure 4, we see that compared to the balanced dataset $\mathcal{D}_{\text{bal}}$, although the full dataset $\mathcal{D}_{\text{full}}$ has strictly more samples with the accurate kind of features, when σ_ξ is not too large, the test accuracy is consistently on par or even lower than training on just the balanced subset. In this case, the views are simply features positioned at different pixels. For very large σ_ξ, the test accuracy of the balanced subset can be low because in such case, the full dataset can learn the dominant view well, but the unbalanced dataset has too few samples to learn any view. The experiment shows that even for architectures such as

convolutional networks, which are believed to have some translation invariance, we should not expect samples from one view to help the learning of other views.

Appendix

We clarify that throughout the appendix $c_1, c_2, \ldots$ denote constants, while C denotes the number of channels in our model (3) and is not a constant, but is a function of d. Throughout the appendix, for any sample $(\mathbf{x}^{(i)}, y^{(i)})$, we let $\mathcal{P}_{bp}^{(i)}$ be the background patches of $(\mathbf{x}^{(i)}, y^{(i)})$ and for $k \in [K]$, $\mathcal{P}_{bp,k}^{(i)}$ be the background patches with feature noise $-\alpha_{p,i} y \mathbf{v}_k$.

B. Useful concentration lemmas

We first state the following standard results on Gaussian samples. These will be used in our proof frequently. .

Lemma 2 (Laurent-Massart χ^2 tail bound)**.** *Consider a standard Gaussian vector $\mathbf{z} \sim \mathcal{N}(0, I_d)$. For any positive vector $\mathbf{a} \in \mathbb{R}^d_{\geq 0}$, and any $t > 0$, the following concentration holds*

$$Pr\Big[\sum_{i=1}^{d} \mathbf{a}_i \mathbf{z}_i^2 \geq \|\mathbf{a}\|_1 + 2\|\mathbf{a}\|_2 \sqrt{t} + 2\|\mathbf{a}\|_\infty t\Big] \leq \exp(-t),$$

$$Pr\Big[\sum_{i=1}^{d} \mathbf{a}_i \mathbf{z}_i^2 \leq \|\mathbf{a}\|_1 - 2\|\mathbf{a}\|_2 \sqrt{t}\Big] \leq \exp(-t).$$

The following corollary immediately follows from using $t = \log(2/\delta)$ and $\mathbf{a}_i = 1$ in the above lemma

Corollary 3 (ℓ_2 norm of Gaussian vector)**.** *Consider $\mathbf{z} \sim \mathcal{N}(0, \sigma^2 I_d)$, for any $\delta \in (0,1)$ and large enough d, we have with probability greater than $1 - \delta$,*

$$\sigma^2 d\left(1 - 2\sqrt{\frac{\log(2/\delta)}{d}}\right) \leq \|\mathbf{z}\|_2^2 \leq \sigma^2 d\left(1 + 4\sqrt{\frac{\log(2/\delta)}{d}}\right).$$

Lemma 4 (Gaussian correlation)**.** *Consider independently sampled Gaussian vectors $\mathbf{z}_1 \sim \mathcal{N}(0, \sigma_1^2 I_d)$ and $\mathbf{z}_2 \in \mathcal{N}(0, \sigma_2^2 I_d)$. For any $\delta \in (0,1)$ and large enough d, there exists a constant c_1, c_2 such that*

$$|\mathbf{z}_1 \cdot \mathbf{z}_2| \leq c_1 \sigma_1 \sigma_2 \sqrt{d \log(2/\delta)} \quad w.p \geq 1 - \delta,$$
$$\mathbf{z}_1 \cdot \mathbf{z}_2 \geq c_2 \sigma_1 \sigma_2 \sqrt{d} \quad w.p \geq 1/4.$$

Proof. Let $u = \|\mathbf{z}_2\|_2$ and $v = \mathbf{z}_1 \cdot \frac{\mathbf{z}_2}{\|\mathbf{z}_2\|_2}$. Since $\mathbf{z}_1$ is spherically symmetric, we have $v \sim \mathcal{N}(0, \sigma_1^2)$ and is independent of u. We first show the upper bound.

$$\begin{aligned}
\Pr(|uv| \geq t) &= \Pr(|v| \geq t/u, u \geq c) + \Pr(|v| \geq t/u, u \leq c) && \text{(holds } \forall\, c > 0) \\
&\leq \min_{c>0} \Pr(u \geq c) + \Pr(|v| \geq t/c) \\
&\leq \Pr\Big(u \geq 2\sigma_2\sqrt{d}\Big) + \Pr\left(|v| \geq \frac{t}{2\sigma_2\sqrt{d}}\right) && \text{(using } c = 2\sigma_2\sqrt{d}) \\
&\leq \exp\left(-\frac{d}{4}\right) + \exp\left(-\frac{t^2}{8\sigma_1^2\sigma_2^2 d}\right) && \text{(using Lemma 2 on } u = \|\mathbf{z}_2\|_2 \text{ and } v \sim \mathcal{N}(0, \sigma_1^2))
\end{aligned}$$

Using $t = 2\sigma_1\sigma_2\sqrt{2d\log(2/\delta)}$, we get the first inequality for all $d \geq 4\log(2/\delta)$.

For the lower bound, using a similar argument as above we have

$$\begin{aligned}
\Pr(uv \leq t) &\leq \min_{c>0} \Pr(v \leq t/c) + \Pr(u \leq c) \\
&\leq \Pr\Big(v \leq \frac{\sigma_1}{4}\Big) + \Pr\left(u \leq \frac{1}{2}\sigma_2\sqrt{d}\right) && \text{(using } c = \tfrac{1}{2}\sigma_2\sqrt{d} \text{ and } t = \tfrac{1}{8}\sigma_1\sigma_2\sqrt{d})
\end{aligned}$$

$$\leq \frac{5}{8} + \exp\left(-\frac{d}{16}\right) \leq \frac{3}{4} \qquad \text{(using Lemma 2 on } u \text{ and cdf bound on } v\text{)}$$

The lower bound thus holds for $d \geq 16\log(8)$ using $t = \frac{1}{8}\sigma_1\sigma_2\sqrt{d}$. This concludes the proof of the lemma. □

Lemma 5 (Gaussian tail concentration)**.** *Consider i.i.d samples* $\{z_c \sim \mathcal{N}(0, \sigma^2) : c \in [C]\}$*. We have the following:*

$$\max_{c\in[C]} |z_c| \leq \sigma\sqrt{2\log\frac{2C}{\delta}}, \quad \textit{w.p} \geq 1-\delta,$$
$$\max_{c\in[C]} z_c \geq \frac{\sigma}{2}, \quad \textit{w.p} \geq 1-\exp(-C/4).$$

Proof. These are standard Gaussian tail bounds, which we prove here for completeness. We have:

$$\Pr\left(\max_{c\in[C]} z_c \geq t\right) \leq \sum_{c\in[C]} \Pr(z_c \geq t) \leq C\exp\left(\frac{-t^2}{2\sigma^2}\right).$$

Using the same argument for over $2C$ variables $\{z_c \sim \mathcal{N}(0,\sigma^2), -z_c \sim \mathcal{N}(0,\sigma^2)\}_{c\in[C]}$ along with $t = \sigma\sqrt{2\log(2C/\delta)}$, we have the first inequality that $\max_{c\in[C]}|z_c| \leq \sigma\sqrt{2\log\frac{2C}{\delta}}$, w.p $\geq 1-\delta$.

Furthermore, $\forall_{c\in[C]}$, we have $\Pr(z_c \geq \sigma/2) \geq 1/4$, hence

$$\Pr\left(\max_{c\in[C]} z_c \geq \sigma/2\right) \geq 1 - \left(1-1/4\right)^C \geq 1-\exp(-C/4)$$

This concludes the proof of the lemma. □

Lemma 8 (Berry–Esseen theorem (Berry, 1941))**.** *Consider i.i.d samples* $\{u_i : i \in [n]\}$ *with* $\mathbb{E}u_i = 0$*,* $\mathbb{E}u_i^2 = \sigma^2 > 0$ *and* $\mathbb{E}|u_i|^3 = \rho < \infty$*. Let* F_n *be the cumulative distribution function of* $\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^n u_i$*, and* Φ *be the cumulative distribution function of the standard normal distribution. For all* t*, there exists a constant* c_1 *such that*

$$|F_n(t) - \Phi(t)| \leq \frac{c_1\rho}{\sigma^3\sqrt{n}}.$$

Lemma 9 (Anti-concentration of q-th power of Gaussian random variables)**.** *Consider i.i.d samples* $\{z_c \sim \mathcal{N}(0,1) : c \in [C]\}$. For constant integer $q \geq 1$, there exist constants $c_1, c_2 > 0$ *such that for any* $t \leq o(1)$*,*

$$\Pr\left[\sum_{c\in[C]} z_c^q \geq c_1 t\sqrt{C}\right] \geq \frac{1}{2} - o(1) - \frac{c_2}{\sqrt{C}}.$$

Proof. For constant q, $\mathbb{E}z_c^{2q} \leq O(1)$ and $\mathbb{E}|z_c|^{3q} \leq O(1)$ (Elandt, 1961). Then, by Lemma 8, for any t, there exist c_1 and c_2 such that

$$\Pr\left[\frac{1}{c_1\sqrt{C}}\sum_{c\in[C]} z_c^q \geq t\right] \geq \Pr\left[z_1 \geq t\right] - \frac{c_2}{\sqrt{C}}.$$

Choosing $t = o(1)$ proves the lemma. □

C. Additional notation

Recall the data distribution $\mathcal{D}$ from Definition 1. Further recall that, for $i \in [n]$, we use k_i^*, p_i^*, p_i^ξ, $\boldsymbol{\xi}^{(i)}$, and $(\alpha_{p,i}, k_{p,i})_{p\notin\{p_i^*, p_i^\xi\}}$ to denote the respective quantities k^*, p^*, p^ξ, $\boldsymbol{\xi}$, and $(\alpha_p, k_p)_{p\notin\{p^*,p^\xi\}}$ in Definition 1 for the i^{th} training sample $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}_{\text{train}} \sim \mathcal{D}$. In addition to these notation in Section 2, we introduce the following additional notation for the proofs.

1. $\forall\, k \in [K]$, let $\mathcal{I}_k = \{i \in [n] : k_i^* = k\}$ denote the set of indices of the training data (x, y) with $y\mathbf{v}_k$ as the main feature. Further, let $n_k = |\mathcal{I}_k|$

2. $\forall\, i \in [n]$ and $\forall\, k \in [K]$, let $\mathcal{P}_{bp,k}^{(i)}$ be the background patches of the i^{th} sample with k^{th}-type feature noise, *i.e.,*

$$\mathcal{P}_{bp,k}^{(i)} = \{p \in [P] \setminus \{p_i^*, p_i^\xi\} : \mathbf{x}_p^{(i)} = -\alpha_{p,i} y \mathbf{v}_k\};$$

and let $\mathcal{P}_{bp}^{(i)} = \bigcup_{k\in[K]} \mathcal{P}_{bp,k}^{(i)} = [P] \setminus \{p_i^*, p_i^\xi\}$ denote the set of all background patches of the i^{th} sample.

Remark 1. *For $k \in [K]$, let $\hat{\rho}_k = \frac{1}{n}|\mathcal{I}_k|$ denote the empirical fraction in the training data of k^{th}. Recall that k_i are sampled independently with $Pr(k_i^* = k) = \rho_k$. Thus, with with high probability, ρ_k and $\hat{\rho}_k$ differ at most by $\sqrt{\frac{\log(n)}{n}}$. In the rest of the paper, for simplicity we assume $\rho_k = \hat{\rho}_k$.*

Similarly, let $\hat{\rho}_k^{(noise)}$ be the proportion of feature noise $-y\mathbf{v}_k$ in dataset $\mathcal{D}_{train}$, i.e., $\hat{\rho}_k^{(noise)} = \frac{1}{n(P-2)}|\{i \in [n], p \in [P] \setminus \{p_i^*, p_i^\xi\}]|k_{p,i} = k\}|$ Again from standard concentration, we have ρ_k and $\hat{\rho}_k^{(noise)}$ differ by negligible quantity with high probability, thus we also assume $\rho_k = \hat{\rho}_k^{(noise)}$.

D. Proof of initialization conditions in Lemma 1

Lemma 1. *[$\mathcal{G}_{\text{init}}$-conditions] Consider n i.i.d. samples $\mathcal{D}_{train} = \{(\mathbf{x}^{(i)}, y^{(i)}) : i \in [n]\}$ from the distribution in Definition 1. Let the parameters $\mathbf{w}$ of the network in (3) be initialized as $\mathbf{w}_c(0) \sim \mathcal{N}(0, \sigma_0^2 I_d)\ \forall\, c \in [C]$. If the number of channels is $C = \Omega(\log d)$, then with probability greater than $1 - O(\frac{n^2 KC}{\text{poly}(d)})$, the following conditions hold :*

1. *Feature-vs-parameter: $\forall\, k \in [K]$, $\max_{c\in[C]} \mathbf{w}_c(0) \cdot \mathbf{v}_k \geq \Omega(\sigma_0)$, and $\max_{c\in[C]} |\mathbf{w}_c(0) \cdot \mathbf{v}_k| \leq \widetilde{O}(\sigma_0)$.*

2. *Noise-vs-parameter:$\forall\, i \in [n]$, $\max_{c\in[C]} \mathbf{w}_c(0) \cdot y^{(i)}\boldsymbol{\xi}^{(i)} \geq \widetilde{\Omega}(\sigma_0\sigma_\xi)$, and $\max_{c\in[C]} |\mathbf{w}_c(0) \cdot \boldsymbol{\xi}^{(i)}| \leq \widetilde{O}(\sigma_0\sigma_\xi)$.*

3. *Noise-vs-noise: $\forall\, i \in [n]$, $\boldsymbol{\xi}^{(i)} \cdot \boldsymbol{\xi}^{(i)} = \Theta(\sigma_\xi^2)$ and $\forall\, i, j \in [n], i \neq j$, $|\boldsymbol{\xi}^{(i)} \cdot \boldsymbol{\xi}^{(j)}| \leq \widetilde{O}(\sigma_\xi^2/\sqrt{d})$.*

4. *Feature-vs-noise: $\forall\, i \in [n], k \in [K]$, $|\boldsymbol{\xi}^{(i)} \cdot \mathbf{v}_k| \leq \widetilde{O}(\sigma_\xi/\sqrt{d})$.*

5. *Parameter norm: $\forall\, c \in [C]$, $\|\mathbf{w}_c(0)\| = \Theta(\sigma_0\sqrt{d})$.*

Proof. Recall the setting of the lemma: $\forall\, k \in [K]$, $\|\mathbf{v}_k\|_2 = 1$, $\forall\, i \in [n]$, $y^{(i)}\boldsymbol{\xi}^{(i)} \overset{i.i.d}{\sim} \mathcal{N}(0, \frac{\sigma_\xi^2}{d} I_d)$, and $\forall\, c \in [C]$, $\mathbf{w}_c(0) \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_0^2 I_d)$. We have the following arguments that prove the lemma, where we use $\delta = \frac{1}{\text{poly}(d)}$.

1. *Feature parameter correlations:* $\forall\, k \in [K]$, we have $\{\mathbf{w}_c(0) \cdot \mathbf{v}_k \sim \mathcal{N}(0, \sigma_0^2)\}_{c\in[C]}$ are C i.i.d Gaussian. Thus, using union bound on the Gaussian tail concentration in Lemma 5 we have condition (1) holds w.p. $\geq 1 - K\delta - K\exp(-C/4)$.

2. *Noise-parameter correlation:* $\forall\, i \in [n]$ and $\forall\, c \in [C]$ using Gaussian correlation bound from Lemma 4, we have $|\mathbf{w}_c(0) \cdot \boldsymbol{\xi}^{(i)}| \geq \widetilde{O}(\sigma_0\sigma_\xi)$ w.p. $\geq 1 - nC\delta$.

   Furthermore, using the second inequality in Lemma 4, we have $\mathbf{w}_c(0) \cdot y^{(i)}\boldsymbol{\xi}^{(i)} \geq c_2\sqrt{\sigma_0\sigma_\xi}$ w.p. $\geq 1/4$. Hence, $\max_{c\in[C]} \mathbf{w}_c(0) \cdot y^{(i)}\boldsymbol{\xi}^{(i)} \geq c_2\sqrt{\sigma_0\sigma_\xi}$ w.p. $\geq 1 - (1 - 1/4)^C \geq 1 - \exp(C)$.

   Thus, summing over failure probabilities, we have that condition (2) holds w.p. $\geq 1 - nC\delta - n\exp(-C/4)$

3. *Noise-noise correlations:* Using the ℓ_2 norm bound from Corollary 3 on $\|\boldsymbol{\xi}^{(i)}\|_2^2$, and the correlation tail bound on $|\boldsymbol{\xi}^{(i)} \cdot \boldsymbol{\xi}^{(j)}|$ for $i \neq j$ from Lemma 4, we have condition (3) holds w.p. $\geq 1 - 2n^2\delta$

4. *Feature noise correlation:* $\forall\, k \in [K]$, we have $\{\boldsymbol{\xi}^{(i)} \cdot \mathbf{v}_k \sim \mathcal{N}(0, \sigma_\xi^2/d)\}_{i\in[n]}$ are n i.i.d Gaussians. Thus, again using union bound on the Gaussian tail concentration in Lemma 5 condition (4) holds w.p. $\geq 1 - n\delta$.

5. *Parameter norm:* From concentration of ℓ_2 norm of Gaussian vector in Corollary 3, condition (5) holds w.p. $\geq 1 - 2C\delta$.

The lemma follows from using $\delta = \frac{1}{\text{poly}(d)}$ and $C = \Omega(\log d) \Rightarrow \exp(-C) = O(\frac{1}{\text{poly}(d)})$. □

Lemma 1a. *$\mathcal{G}_{\text{init}}$ in Lemma 1 also holds for $\mathcal{D}^{(aug)}_{train}$ defined in (2) with n replaced by nK.*

Proof. Recall that since the features $\{\mathbf{v}_k\}_k$ are orthonormal (Assumption 1) and all the non-feature noise are spherically symmetric, without loss of generality, we assume that $\{\mathbf{v}_k\}_{k\in[K]}$ are simply the first K standard basis vectors in $\mathbb{R}^d$, *i.e.,* $\mathbf{v}_k = \mathbf{e}_k$. In this case, we choose $\mathcal{T}_k$ for $k \in [K-1]$ as a permutation of coordinates of $\mathbb{R}^d$ without any fixed points, *i.e.,* $\forall\, i \in [d]$, $\mathcal{T}_k(\mathbf{z})[i] \neq \mathbf{z}[i]$ that satisfies (1) on the first K coordinate.

We now show that the $\mathcal{G}_{\text{init}}$ conditions in Lemma 1 holds for $\mathcal{D}^{(\text{aug})}_{\text{train}} = \mathcal{D}_{\text{train}} \cup \mathcal{T}_1(\mathcal{D}_{\text{train}}) \cup \mathcal{T}_2(\mathcal{D}_{\text{train}}) \cup \ldots \cup \mathcal{T}_{K-1}(\mathcal{D}_{\text{train}})$ defined with transformations $\{\mathcal{T}_k\}_{k\in[K-1]}$ described above.

- First, among the $\mathcal{G}_{\text{init}}$ conditions, (1) and (5) are independent of the samples and hence immediately hold.
- Secondly, $\forall\, i \in [n]$ and $\forall\, k \in [K]$, $\mathcal{T}_k(\boldsymbol{\xi}^{(i)})$ is simply some permutation of the coordinates of $\boldsymbol{\xi}^{(i)} \sim \mathcal{N}(0, \sigma_\xi^2 I_d/d)$, and hence $\mathcal{T}_k(\boldsymbol{\xi}^{(i)}) \sim \mathcal{N}(0, \sigma_\xi^2 I_d/d)$ has the same marginal distribution as $\boldsymbol{\xi}^{(i)}$. This implies that conditions (2) and (4), as well the norm condition in (3) of Lemma 1 also holds for $\mathcal{D}^{(\text{aug})}_{\text{train}}$.
- Finally, note that $\forall\, i \neq j$, $\forall\, k, k'$, $\mathcal{T}_k(\boldsymbol{\xi}^{(i)})$ and $\mathcal{T}_{k'}(\boldsymbol{\xi}^{(j)})$ are independent Gaussians. Thus, the correlation bounds in (3) of the form $|\mathcal{T}_k(\boldsymbol{\xi}^{(i)}) \cdot \mathcal{T}_{k'}(\boldsymbol{\xi}^{(j)})| = \widetilde{O}(\sigma_\xi^2/\sqrt{d})$ for all $i \neq j$ also follow from the proof of Lemma 1.

The only non-trivial condition we want to show is the following bound on the noise correlations of distinct transformations of the same sample, *i.e.,* we only need to show that $|\boldsymbol{\xi}^{(i)} \cdot \mathcal{T}_k(\boldsymbol{\xi}^{(i)})| \leq \widetilde{O}(\sigma_\xi^2/\sqrt{d})$ with high probability for all $k \in [K-1]$. Note that for any $1 \leq k < k' \leq K-1$, $\mathcal{T}_k(\boldsymbol{\xi}^{(i)}) \cdot \mathcal{T}_{k'}(\boldsymbol{\xi}^{(i)})$ is equivalent in distribution to $\boldsymbol{\xi}^{(i)} \cdot \mathcal{T}_{k'-k}(\boldsymbol{\xi}^{(i)})$.

Claim 1. *If $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_\xi^2 I_d/d)$ then $\forall\, k \in [K-1]$, $|\boldsymbol{\xi} \cdot \mathcal{T}_k(\boldsymbol{\xi})| \leq O\left(\sigma_\xi^2 \sqrt{\frac{\log(1/\delta)}{d}}\right)$ w.p. $\geq 1 - \delta$.*

Proof. At a high level, we create a non-overlapping partition of the entries of $\boldsymbol{\xi}$ into three vectors $\boldsymbol{\xi}'$, $\boldsymbol{\xi}''$, and $\boldsymbol{\xi}'''$, each of which of length at least $d/6$. The partition is chosen such that same partitioning of entries of $\mathcal{T}_k(\boldsymbol{\xi})$ denoted as $\tilde{\boldsymbol{\xi}}'$, $\tilde{\boldsymbol{\xi}}''$, and $\tilde{\boldsymbol{\xi}}'''$ are independent of $\boldsymbol{\xi}'$, $\boldsymbol{\xi}''$, and $\boldsymbol{\xi}'''$, respectively. We then have $\boldsymbol{\xi} \cdot \mathcal{T}_k(\boldsymbol{\xi}) = \boldsymbol{\xi}' \cdot \tilde{\boldsymbol{\xi}}' + \boldsymbol{\xi}'' \cdot \tilde{\boldsymbol{\xi}}'' + \boldsymbol{\xi}''' \cdot \tilde{\boldsymbol{\xi}}'''$, where each term is a dot product of two independent spherical Gaussians of length at least $d/6$ and entrywise variance of σ_ξ^2/d. The claim then follows from bounding each term using Lemma 4.

We divide the coordinates of $\boldsymbol{\xi}$ into disjoint and ordered lists $L_1, L_2, \ldots$, constructed as follows. The first list is

$$L_1 = \left[\boldsymbol{\xi}[1],\, \mathcal{T}_k(\boldsymbol{\xi})[1],\, \mathcal{T}_k^2(\boldsymbol{\xi})[1], \ldots, \mathcal{T}_k^{s_1}(\boldsymbol{\xi})[1]\right],$$

where $\mathcal{T}_k^m$ denotes composition of $\mathcal{T}_k$ for m times, and we stop the list at the first $s_1 \leq d-1$ such that $\mathcal{T}_k^{s_1+1}(e_1) = e_1$ (when $\mathcal{T}_k^{s_1+1}(\boldsymbol{\xi})[1] = \boldsymbol{\xi}[1]$). We claim that this stopping criteria ensures that L_1 has s_1 unique coordinate of $\boldsymbol{\xi}$ without any duplicates. If not, there exists some $0 \leq s' < s'' \leq s_1$ such that $\mathcal{T}_k^{s''}(e_1) = \mathcal{T}_k^{s'}(e_1)$. Since $\mathcal{T}_k$ is a permutation (hence invertible), this would imply that $\mathcal{T}_k^{s''-s'}(e_1) = e_1$ for $s'' - s' \leq s_1$, which contradicts the stopping criteria.

Note that if $s_1 = d-1$, we have included all the coordinates of $\boldsymbol{\xi}$ in L_1, and we stop our stop our construction here. If L_1 does not contain all coordinates of $\boldsymbol{\xi}$, let $1 < j_2 \leq d$ be the first coordinate such that $\boldsymbol{\xi}[j_2] \notin L_1$. Let,

$$L_2 = \left[\boldsymbol{\xi}[j_2],\, \mathcal{T}_k(\boldsymbol{\xi})[j_2],\, \mathcal{T}_k^2(\boldsymbol{\xi})[j_2], \ldots, \mathcal{T}_k^{s_2}(\boldsymbol{\xi})[j_2]\right],$$

where we stop either when all the entries of $\boldsymbol{\xi}$ have been included in $L_1^{(m)}$ or $L_2^{(m)}$, or at the first integer s_2 such that $\mathcal{T}_k^{s_2+1}(e_{j_2}) = e_{j_2}$ (when $\mathcal{T}_k^{s_2+1}(\boldsymbol{\xi})[j_2] = \boldsymbol{\xi}[j_2]$). With a similar argument as with L_1, there are no duplicate coordinates in L_2. Furthermore, we either have have L_2 and L_1 containing disjoint coordinates of $\boldsymbol{\xi}$, or have $L_1 \subset L_2$. To see this, suppose for $0 \leq s' \leq s_1$ and $0 \leq s'' \leq s_2$, we have $\mathcal{T}_k^{s'}(e_1) = \mathcal{T}_k^{s''}(e_{j_2})$. If $s' \geq s''$, again from invertibility of $\mathcal{T}_k$, we

would have $\mathcal{T}_k^{s'-s''}(e_1) = e_{j_2}$ for $s' - s'' \leq s_1$, which is contradiction for $\boldsymbol{\xi}[j_2] \notin L_1$. On the other hand, if $s' < s''$, then $\mathcal{T}_k^{s''-s'}(e_{j_2}) = e_1$, and the entire construction of L_1 would also be contained in L_2. This would imply that all the coordinates of L_1 are contained in L_2 exactly once (since L_2 does not have duplicates). Without loss of generality, we assume the former condition that L_2 and L_1 are disjoint holds as otherwise, $L_1 \subset L_2$ and we can simply drop the first list L_1 from our construction, and our proof follows exactly.

We construct $L_3, L_4, \ldots, L_\ell$ similarly until all coordinates of $\boldsymbol{\xi}$ belong to exactly one list. We also define $\mathcal{T}L_1, \mathcal{T}L_2, \ldots, \mathcal{T}L_\ell$ as lists obtained by circularly shifting the coordinates of $L_1, L_2, \ldots, L_\ell$, respectively, by one index. For example, $\mathcal{T}L_1 = \big[\mathcal{T}_k(\boldsymbol{\xi})[1], \mathcal{T}_k^2(\boldsymbol{\xi})[1], \ldots, \mathcal{T}_k^{s_1}(\boldsymbol{\xi})[1],\, \boldsymbol{\xi}[1]\big]$.

By construction, for $l = 1, 2 \ldots \ell$, for every coordinate of $\boldsymbol{\xi}$ that is included in L_l, has the same coordinate of $\mathcal{T}_k(\boldsymbol{\xi})$ is included in $\mathcal{T}L_l$ at the same position, *i.e.,* for all $i \leq s_l, j \leq d$, $L_l[i] = \boldsymbol{\xi}[j] \implies \mathcal{T}L_l[i] = \mathcal{T}(\boldsymbol{\xi})[j]$. We now construct $\boldsymbol{\xi}'$, $\boldsymbol{\xi}''$, and $\boldsymbol{\xi}'''$. For $l = 1, 2 \ldots, \ell$, do the following:

- Sequentially distribute all the elements *except the last element* of L_l to $\boldsymbol{\xi}', \boldsymbol{\xi}'', \boldsymbol{\xi}'''$, *e.g.,* the 1st element of L_l goes to $\boldsymbol{\xi}'$, 2nd to $\boldsymbol{\xi}''$, 3rd to $\boldsymbol{\xi}'''$, 4th to $\boldsymbol{\xi}'$ and so on. This assignment ensures that $\boldsymbol{\xi}', \boldsymbol{\xi}'', \boldsymbol{\xi}'''$ do not contain any adjacent entries of L_l, *i.e.,* if $L_l[i]$ is in $\boldsymbol{\xi}'$, then $L_l[i+1]$ is not in $\boldsymbol{\xi}'$, and same is true for $\boldsymbol{\xi}''$, and $\boldsymbol{\xi}''$.
- Include the last element of L_l to a list among $\boldsymbol{\xi}', \boldsymbol{\xi}'', \boldsymbol{\xi}'''$ that *does not* contain the first or the second last element of L_l. Thus the last element of L_l is not in the same list as its circularly adjacent neighbors $\boldsymbol{\xi}[j_l]$ and $\mathcal{T}_k^{s_l-1}(\boldsymbol{\xi})[j_l]$.
- Repeat the exact assignment as above to distribute the elements of $\mathcal{T}L_l$ to $\tilde{\boldsymbol{\xi}}', \tilde{\boldsymbol{\xi}}^{(\prime\prime}, \tilde{\boldsymbol{\xi}}'''$.

By construction, $\{\boldsymbol{\xi}', \boldsymbol{\xi}'', \boldsymbol{\xi}'''\}$ and $\{\tilde{\boldsymbol{\xi}}', \boldsymbol{\xi}'', \boldsymbol{\xi}'''\}$ satisfy the following properties: (a) $\boldsymbol{\xi} \cdot \mathcal{T}_k(\boldsymbol{\xi}) = \boldsymbol{\xi}' \cdot \tilde{\boldsymbol{\xi}}' + \boldsymbol{\xi}'' \cdot \tilde{\boldsymbol{\xi}}'' + \boldsymbol{\xi}''' \cdot \tilde{\boldsymbol{\xi}}'''$. (b) $\boldsymbol{\xi}', \boldsymbol{\xi}''$, and $\boldsymbol{\xi}'''$ are independent of $\tilde{\boldsymbol{\xi}}', \tilde{\boldsymbol{\xi}}''$, and $\tilde{\boldsymbol{\xi}}'''$, respectively. Furthermore, each of these vectors is a spherical Gaussian with entrywise variance of σ_ξ^2/d. (c) we have included at least $d/3 - 1 = \Theta(d)$ entries of $\boldsymbol{\xi}$ in each of $\boldsymbol{\xi}', \boldsymbol{\xi}''$, and $\boldsymbol{\xi}'''$. The claim now follows from using Lemma 4 on $\boldsymbol{\xi}' \cdot \tilde{\boldsymbol{\xi}}'$, $\boldsymbol{\xi}'' \cdot \tilde{\boldsymbol{\xi}}''$, and $\boldsymbol{\xi}''' \cdot \tilde{\boldsymbol{\xi}}'''$.

□

The above claim completes the proof of Lemma 1a. □

E. Linear models

In this section we discuss the behavior of linear models for data from our distribution $\mathcal{D}$ in Definition 1. We consider the same patchwise convolutional model in (3), *but without non-linearity*. Without loss of generality, assume $C = 1$. Thus, for $\boldsymbol{\theta} \in \mathbb{R}^d$, the model effectively becomes $f^{\text{linear}}(\boldsymbol{\theta}, \mathbf{x}) = \boldsymbol{\theta} \cdot \bar{\mathbf{x}}$, where $\bar{\mathbf{x}} = \sum_p \mathbf{x}_p$.

Linear models without feature noise. In the first result stated and proved below, we assume no feature noise $\alpha_p = 0$. In this case, $\bar{\mathbf{x}}^{(i)} = y^{(i)}\mathbf{v}_{k_i^*} + \boldsymbol{\xi}^{(i)}$. Recall the notation that for $k \in [K]$, $\mathcal{I}_k = \{i \in [n] : k_i^* = k\}$ and $n_k = |\mathcal{I}_k|$.

Theorem 6. *With high probability, the max ℓ_2 margin linear model over $\mathcal{D}_{train} = \{(\bar{\mathbf{x}}^{(i)}, y^{(i)}) : i \in [n]\}$ is given by*

$$\widehat{\boldsymbol{\theta}}_{\ell_2} = \sum_{k \in [K]} \frac{1}{1 + (1 + o(1))\sigma_\xi^2/n_k} \left(\mathbf{v}_k + \frac{1}{n_k} \sum_{i \in I_k} y^{(i)} \boldsymbol{\xi}^{(i)} \right) \tag{8}$$

Proof. Without loss of generality, assume the data points are grouped by the feature type k_i^*, such that $\mathcal{I}_1 = \{1, 2, \ldots, n_1\}$, $\mathcal{I}_2 = \{n_1 + 1, n_1 + 2, \ldots n_1 + n_2\}$, and so on. Also let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote a matrix containing $y^{(i)}\bar{\mathbf{x}}^{(i)}$ as rows and let $\mathcal{K} = \mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{n \times n}$ denote the corresponding kernel matrix.

The ℓ_2 max margin classifier is given by $\widehat{\boldsymbol{\theta}}_{\ell_2} = \min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_2^2$ s.t. $\mathbf{X}\boldsymbol{\theta} \geq 1$. From the optimality conditions of the max-margin problem, we know that there exists a dual variable $\boldsymbol{\nu} \in \mathbb{R}_+^n$, s.t. $\widehat{\boldsymbol{\theta}}_{\ell_2} = \mathbf{X}^\top \boldsymbol{\nu}$. We use notation $\boldsymbol{\nu}_k \in \mathbb{R}_+^{n_k}$ such that $\boldsymbol{\nu} = [\boldsymbol{\nu}_1^\top, \boldsymbol{\nu}_2^\top, \ldots \boldsymbol{\nu}_K^\top]^\top$. We can now write the objective and constraints of the max margin problem in terms of dual variables as follows: $\|\boldsymbol{\theta}\|_2^2 = \boldsymbol{\nu}^\top \mathcal{K} \boldsymbol{\nu}$ and the margin condition is $\mathcal{K}\boldsymbol{\nu} \geq 1$.

Let us first look at structure of $\mathcal{K}$. Recall that $\bar{\mathbf{x}}^{(i)} = y^{(i)}\mathbf{v}_{k_i^*} + \boldsymbol{\xi}^{(i)}$, where $\{\mathbf{v}_k\}_k$ are orthonormal and $\boldsymbol{\xi}^{(i)} \sim \mathcal{N}(0, \sigma_\xi^2 I_d/d)$. Using the standard concentration inequalities in Appendix B, the following holds with high probability.

$$\mathcal{K}_{ij} = y^{(i)}\bar{\mathbf{x}}^{(i)} \cdot y^{(j)}\bar{\mathbf{x}}^{(j)} = \begin{cases} 1 + \sigma_\xi^2 + \widetilde{O}(\frac{\sigma_\xi^2}{\sqrt{d}}) & \text{if } i = j \\ 1 + \widetilde{O}(\frac{\sigma_\xi^2}{\sqrt{d}}) & \text{if } i \neq j,\ k_i^* = k_j^* \\ \widetilde{O}(\frac{\sigma_\xi^2}{\sqrt{d}}) & \text{if } i \neq j,\ k_i^* \neq k_j^* \end{cases}.$$

We can combine all the $\widetilde{O}(\frac{\sigma_\xi^2}{\sqrt{d}})$ terms in Δ, and write $\mathcal{K} = \bar{\mathcal{K}} + \Delta$ where $\bar{\mathcal{K}}_{ij} = \mathbf{1}_{k_i^*=k_j^*} + \sigma_\xi^2 \mathbf{1}_{i=j}$. Thus, $\bar{\mathcal{K}}$ is a block diagonal matrix which is dominant compared to lower order terms in Δ.

Based on this block dominant structure of $\mathcal{K}$, for $\mathbf{w} = \mathbf{X}^\top \boldsymbol{\nu}$ and $\boldsymbol{\nu} \geq 0$, the margin on data points is given by,

$$\forall\, i \in \mathcal{I}_k,\ (\mathcal{K}\boldsymbol{\nu})_i = \|\boldsymbol{\nu}_k\|_1 + \sigma_\xi^2 \boldsymbol{\nu}_{k,i} + (\Delta\boldsymbol{\nu})_i, \tag{9}$$

and the ℓ_2 norm is given by,

$$\|\boldsymbol{\theta}\|_2^2 = \boldsymbol{\nu}^\top \mathcal{K}\boldsymbol{\nu} = \left(\sum_{k\in[K]} \|\boldsymbol{\nu}_k\|_1^2 + \sigma_\xi^2 \|\boldsymbol{\nu}_k\|_2^2\right) + \boldsymbol{\nu}^\top \Delta\nu. \tag{10}$$

Recall that $\Delta_{ij} = \widetilde{O}(\sigma_\xi^2/\sqrt{d})$, we have the following two possibilities of ν:

Case 1. $\|\boldsymbol{\nu}\|_\infty = O(1)$: In this case $(\Delta\boldsymbol{\nu})_i = o(\sigma_\xi^2)$ and we have $(\mathcal{K}\boldsymbol{\nu})_i = \|\boldsymbol{\nu}_k\|_1 + \sigma_\xi^2\boldsymbol{\nu}_{k,i} + o(\sigma_\xi^2)$. Thus the margin constraint requires that $\min_k \min_{i\in\mathcal{I}_k} \|\boldsymbol{\nu}_k\|_1 + \sigma_\xi^2\boldsymbol{\nu}_{k,i} + o(\sigma_\xi^2) \geq 1$. Furthermore, for large enough d, $\|\boldsymbol{\theta}\|_2^2$ is monotonic in $\boldsymbol{\nu}_{k,i}$ (for positive $\boldsymbol{\nu}_{k,i}$). Thus the optimal $\boldsymbol{\nu}$ is given by

$$\forall\, k \in [K],\ \forall\, i \in \mathcal{I}_k,\ \boldsymbol{\nu}_{k,i} = \frac{1}{n_k + (1+o(1))\sigma_\xi^2}. \tag{11}$$

In this case, $\|\boldsymbol{\theta}\|_2^2 = \frac{1}{1+\sigma_\xi^2/n_k}(1+o(1)) = O(1)$.

Case 2. If $\boldsymbol{\nu} = \omega(1)$, then $\|\boldsymbol{\theta}\|_2^2 = \omega(1)$ which is suboptimal compared to the above case.

In conclusion, we have the optimal $\boldsymbol{\nu}$ for the max-margin problem given by (11). Thus,

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{\ell_2} = \mathbf{X}^\top\boldsymbol{\nu} &= \sum_{k\in[K]}\sum_{i\in I_k} \nu_{k,i} y^{(i)}\bar{\mathbf{x}}^{(i)} \\ &= \sum_{k\in[K]}\sum_{i\in I_k} \frac{\mathbf{v}_k + y^{(i)}\boldsymbol{\xi}^{(i)}}{n_k + (1+o(1))\sigma_\xi^2} \\ &= \sum_{k\in[K]} \frac{1}{1 + (1+o(1))\sigma_\xi^2/n_k}\left(\mathbf{v}_k + \frac{1}{n_k}\sum_{i\in I_k} y^{(i)}\boldsymbol{\xi}^{(i)}\right). \end{aligned}$$

This concludes the proof of the theorem. □

For the above classifier, for simplicity, we look at the case when there are only two views, $k = 2$. Corollary 7 follows from direct calculation on $\widehat{\boldsymbol{\theta}}_{\ell_2}^\top \mathbf{x}$ for a sample $\mathbf{x}$ from our distribution. The thresholds given in Corollary 7 are better than the threshold we derive for our neural network.

Corollary 7. *Suppose $k = 2$, $\omega(1) \leq \sigma_\xi^2 \leq \sqrt{nd}$ and $n \leq d$. The ℓ_2 max-margin linear model in (8) can successfully learn feature $\mathbf{v}_1$. To successfully learn feature $\mathbf{v}_2$, we need $\rho_2 \gg \frac{\sigma_\xi^2}{\sqrt{nd}}$ if $n \leq o(\sigma_\xi^2)$ and $\rho_2 \gg \frac{\sigma_\xi^3}{n\sqrt{d}}$ otherwise.*

Linear models with feature noise. In the second result, we study linear models in the presence of feature noise. We show linear models are not able to learn samples from our data distribution $\mathcal{D}$ while the non-linear model we study can learn $\mathcal{D}$. To facilitate the proof of linear models, we make some additional simplifications. These simplifications are not necessary for our main results. For linear model results alone, we consider the case when the dominant noise $\boldsymbol{\xi}$ is zero, *i.e.,* $\sigma_\xi = 0$. Note that letting $\sigma_\xi > 0$ can only make the classification harder. Let $\Lambda(\mathbf{x})$ be the sum of the coefficients of the feature noise if $\mathbf{x}$, *i.e.,* $\Lambda(\mathbf{x}) = \sum_{k\in[K]} \sum_{p\in\mathcal{P}_{bp}} \alpha_p$. Let μ_Λ be the probability that $\Lambda(\mathbf{x}) > 1$ for each $(\mathbf{x}, y)$. We assume that the patch with the main feature is chosen uniform randomly from $[P]$. Let $\mathcal{D}'$ be the distribution satisfies the above assumptions.

Theorem 10. *For any linear classifier $\boldsymbol{\theta} \in \mathbb{R}^{d\times P}$, we have*

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}'} [\text{sign } \langle \mathbf{x}, \boldsymbol{\theta}\rangle \neq \text{sign } y] > \frac{1}{P} \min\{\mu_\Lambda, 1-\mu_\Lambda\} \min_{k\in[K]} \rho_k.$$

Moreover, there exists a non-linear model F in (3) with weights $\mathbf{w}$, such that

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}'} [\text{sign } F(\mathbf{w}, \mathbf{x}) \neq \text{sign } y] = 0.$$

Proof. Let $\Delta = \min_{p\in[P],k\in[K]} \boldsymbol{\theta}_{(p-1)d+k-1}$ and $p^*, k^* = \arg\min_{p\in[P],k\in[K]} \boldsymbol{\theta}_{(p-1)d+k-1}$. If $\Delta \leq 0$, for any sample with main feature $y\mathbf{v}_{k^*}$ in patch p^*, and $\Lambda(\mathbf{x}) \leq 1$,

$$y\langle \mathbf{x}, \boldsymbol{\theta}\rangle \leq -\Delta + \Lambda(\mathbf{x})\Delta < 0.$$

If $\Delta > 0$, then for any sample with main feature $y\mathbf{v}_{k^*}$ in patch p^*, with $\Lambda(\mathbf{x}) > 1$,

$$y\langle \mathbf{x}, \boldsymbol{\theta}\rangle \leq \Delta - \Lambda(\mathbf{x})\Delta \leq 0.$$

Then, for both the case that $\Delta > 0$ and the case that $\Delta \leq 0$, with probability at least $\min\{\mu_\Lambda, 1-\mu_\Lambda\} \min_{k\in[K]} \rho_k/P$, sign $\langle \mathbf{x}, \boldsymbol{\theta}\rangle \neq$ sign y.

Now, consider the non-linear model given by weights $\mathbf{w}_1 = \sum_{k\in[K]} \mathbf{v}_k$ and $\mathbf{w}_c = 0$ for all $c \in [C]\backslash\{1\}$. For any datapoint $(\mathbf{x}, y)$ with main feature $y\mathbf{v}_{k^*}$,

$$\begin{aligned} yF(\mathbf{w}, \mathbf{x}) &= y \sum_{c\in[C]} \sum_{p\in[P]} \psi(\langle \mathbf{w}_c, \mathbf{x}_p\rangle) \\ &= \psi(\langle \mathbf{w}_1, \mathbf{v}_{k^*}\rangle) - \sum_{k\in[K]} \sum_{\mathcal{P}_{bp,k}} \psi(\langle \mathbf{w}_1, \alpha\mathbf{v}_k\rangle) \\ &\geq \frac{1}{q} - \frac{1}{q}\alpha^q P \\ &> 0. \end{aligned}$$

Thus, we have sign $F(\mathbf{w}, \mathbf{x}) =$ sign y for all samples $(\mathbf{x}, y)$. □

F. Proof of the Main Results

F.1. Dynamics of network weights: learning features and noise

We first present a few lemmas useful for the proof of the main results. We derive the training trajectories for the dataset without data augmentation $\mathcal{D}_{\text{train}}$. All lemmas in this section also hold for the dataset with data augmentation $\mathcal{D}_{\text{train}}^{(\text{aug})}$ with n replaced Kn and ρ_k replaced by $\rho_k^{(\text{aug})} = \frac{1}{K}$. We defer the proof of the lemmas to Appendix G.

Lemma 11 and Lemma 12 give some rough bounds on $\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle$ and $\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\rangle$, which are used repeatedly in the proof.

Lemma 11 (Rough upper and lower bound on $\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle$)**.** *Suppose $\mathcal{G}_{\text{init}}$ holds and*

$$\alpha \leq o\left(\sigma_\xi^{\frac{1}{q}} d^{-\frac{1}{2q}} P^{-\frac{1}{q}} \left(\sigma_0 + \eta T\rho_k + \eta T\sigma_\xi d^{-1/2}\right)^{-\frac{q-1}{q}}\right).$$

For all $0 \le t' \le t \le T$ and $k \in [K]$, we have

$$\max_{c\in[C]} \langle \mathbf{w}_c(t), \mathbf{v}_k \rangle \le \max_{c\in[C]} \langle \mathbf{w}_c(t'), \mathbf{v}_k \rangle + \eta(t-t')\widetilde{O}\left(\rho_k + \sigma_\xi d^{-1/2}\right)$$
$$\le \widetilde{O}\left(\sigma_0 + \eta T\left(\rho_k + \sigma_\xi d^{-1/2}\right)\right),$$

and

$$\min_{c\in[C]} \langle \mathbf{w}_c(t), \mathbf{v}_k \rangle \ge \min_{c\in[C]} \langle \mathbf{w}_c(t'), \mathbf{v}_k \rangle - \eta(t-t')\widetilde{O}\left(\sigma_\xi d^{-1/2}\right)$$
$$\ge -\widetilde{O}\left(\sigma_0 + \eta T \sigma_\xi d^{-1/2}\right).$$

Lemma 12 (Rough lower bound on $\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \rangle$)**.** *Suppose $\mathcal{G}_{\text{init}}$ holds and*

$$\alpha \le \widetilde{O}\left(\min\left\{1, \sigma_\xi^{\frac{1}{q}} d^{-\frac{1}{2q}}\right\} P^{-1/q}\left(\sigma_0 + \eta T\left(\max_{k\in[K]} \rho_k + \sigma_\xi d^{-1/2}\right)\right)^{-(q-1)/q}\right).$$

For all $0 \le t \le t' \le T$ and $i \in [n]$, we have

$$\min_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle \ge \min_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t'), \boldsymbol{\xi}^{(i)} \right\rangle - \eta(t-t')\widetilde{O}\left(\sigma_\xi^2 d^{-1/2} + \sigma_\xi d^{-1/2}\right).$$

Combining Lemma 11 and Lemma 12, we can show that when the time step T is bounded, $\langle \mathbf{w}_c(t), \mathbf{v}_k \rangle$ and $\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \rangle$ are lower bounded.

Lemma 13 (Lower bound on $\langle \mathbf{w}_c(t), \mathbf{v}_k \rangle$ and $\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \rangle$)**.** *Suppose $\mathcal{G}_{\text{init}}$ holds,* $n \le o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^q d^{1/2}, \sigma_0^{q-1}\sigma_\xi^{q-1} d^{1/2}\right\}\right)$, $K \le o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^{-1} d^{1/2}, \sigma_0^{q-1} d^{1/2}\right\}\right)$, *and*

$$\alpha \le \widetilde{O}\left(\min\left\{1, \sigma_\xi^{\frac{1}{q}} d^{-\frac{1}{2q}}\right\} P^{-1/q}\left(\sigma_0 + \eta T\left(\max_{k\in[K]} \rho_k + \sigma_\xi d^{-1/2}\right)\right)^{-(q-1)/q}\right).$$

for some $T = \widetilde{\Theta}\left(\max\left\{n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}, K\eta^{-1}\sigma_0^{-q+2}\right\}\right)$. For all $0 \le t' \le t \le T$, and $c \in [C]$,

$$\langle \mathbf{w}_c(t), \mathbf{v}_k \rangle \ge \langle \mathbf{w}_c(t'), \mathbf{v}_k \rangle - o(\sigma_0),$$

and for all $i \in [n]$,

$$y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle \ge y^{(i)}\left\langle \mathbf{w}_c(t'), \boldsymbol{\xi}^{(i)} \right\rangle - o\left(\sigma_0\sigma_\xi\right).$$

Next, Lemma 14 and Lemma 15 compute the time it takes for the model to learn the main feature $\mathbf{v}_k$, $k \in [K]$ and overfit the noise $\boldsymbol{\xi}^{(i)}$, $i \in [n]$. Lemma 16 and Lemma 17 upper bound $\langle \mathbf{w}_c(t), \mathbf{v}_k \rangle$ and $\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \rangle$ for t smaller than the time identified in Lemma 14 and Lemma 15.

Lemma 14 (Learning the main feature)**.** *Suppose $\mathcal{G}_{\text{init}}$ holds, $C = \Theta(\log d)$, $\sigma_0\sigma_\xi \le o(1)$, $\sigma_\xi^q d^{-1/2} \le o(\rho_k)$ and*

$$\alpha \le o\left(P^{-\frac{1}{q}} \min\left\{1, \sigma_\xi^{\frac{1}{q}} d^{-\frac{1}{2q}}\left(\sigma_0 + \eta T \max_{k\in[K]} \rho_k + \eta T \sigma_\xi d^{-1/2}\right)^{-\frac{q-1}{q}}, \left(\sigma_0 + \eta T \max_{k\in[K]} \rho_k + \eta T \sigma_\xi d^{-1/2}\right)^{-1}\right\}\right),$$

for some $T \ge \widetilde{\Omega}\left(\left(\eta\rho_k\sigma_0^{q-2}\right)^{-1}\right)$. For any $k \in [K]$ and $0 \le t \le T$, if

$$\max_{c\in[C]} \langle \mathbf{w}_c(t), \mathbf{v}_k \rangle \le O(C^{-1/q}), \quad \textit{and} \quad \max_{i\in[n], c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle \le \widetilde{O}(\sigma_0\sigma_\xi),$$

then

$$\max_{c\in[C]} \langle \mathbf{w}_c(t+1), \mathbf{v}_k \rangle = \max_{c\in[C]} \langle \mathbf{w}_c(t), \mathbf{v}_k \rangle + \Theta\left(\eta\rho_k \psi'\left(\max_{c\in[C]} \langle \mathbf{w}_c(t), \mathbf{v}_k \rangle\right)\right).$$

Moreover, if $\max_{i\in[n],c\in[C]} \left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle \le \widetilde{O}(\sigma_0\sigma_\xi)$ *for all* $t \le \widetilde{O}\left(\frac{1}{\eta\rho_k\sigma_0^{q-2}}\right)$*, there exists* $T' \le \widetilde{O}\left(\frac{1}{\eta\rho_k\sigma_0^{q-2}}\right)$ *such that* $\max_{c\in[C]} \langle \mathbf{w}_c(T'), \mathbf{v}_k \rangle \ge \Omega\left(C^{-1/q}\right)$.

Lemma 15 (Overfitting the dominant noise)**.** *Suppose* $\mathcal{G}_{\text{init}}$ *holds,* $C = \Theta(\log d)$*,* $n \le o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^q d^{1/2}, \sigma_0^{q-1}\sigma_\xi^{q-1} d^{1/2}\right\}\right)$ *and*

$$\alpha \le o\left(P^{-\frac{1}{q}} \min\left\{1, \sigma_\xi^{\frac{1}{q}} d^{-\frac{1}{2q}} \left(\sigma_0 + \eta T \max_{k\in[K]} \rho_k + \eta T \sigma_\xi d^{-1/2}\right)^{-\frac{q-1}{q}}, \left(\sigma_0 + \eta T \max_{k\in[K]} \rho_k + \eta T \sigma_\xi d^{-1/2}\right)^{-1}\right\}\right),$$

for some $T \ge \widetilde{\Omega}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$.

Let $i \in [n]$ *be some sample such that for all* $0 \le t \le T$, $\max_{c\in[C]} \left\langle \mathbf{w}_c(t), \mathbf{v}_{k_i^*} \right\rangle \le O(C^{-1/q})$*. For any time step* $0 \le t \le T$ *, if*

$$\max_{c\in[C]} y^{(i)} \left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle \le O(C^{-1/q}),$$

we have

$$\max_{c\in[C]} y^{(i)} \left\langle \mathbf{w}_c(t+1), \boldsymbol{\xi}^{(i)} \right\rangle = \max_{c\in[C]} y^{(i)} \left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle + \frac{\eta}{n}\widetilde{\Theta}\left(\sigma_\xi^2 \psi'\left(\max_{c\in[C]} y^{(i)} \left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle\right)\right).$$

Moreover, there exists times step $T' \le \widetilde{O}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$ *such that* $\max_{c\in[C]} y^{(i)} \left\langle \mathbf{w}_c(T'), \boldsymbol{\xi}^{(i)} \right\rangle \ge \Omega\left(C^{-1/q}\right)$.

Lemma 16 (Upper bound on $\langle \mathbf{w}_c(t), \mathbf{v}_k \rangle$)**.** *If* $\mathcal{G}_{\text{init}}$ *holds, for all* $k \in [K]$ *and* $t \le o\left(\frac{\sigma_0}{\eta\rho_k\sigma_0^{q-1} + \eta\sigma_\xi d^{-1/2}}\right)$*,*

$$\max_{c\in[C]} \langle \mathbf{w}_c(t), \mathbf{v}_k \rangle \le \widetilde{O}\left(\sigma_0\right).$$

Lemma 17 (Upper bound on $\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \rangle$)**.** *Suppose* $\mathcal{G}_{\text{init}}$ *holds,* $n \le o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^q d^{1/2}, \sigma_0^{q-1}\sigma_\xi^{q-1} d^{1/2}\right\}\right)$ *and*

$$\alpha \le o\left(P^{-\frac{1}{q}} \min\left\{1, \sigma_\xi^{\frac{1}{q}} d^{-\frac{1}{2q}} \left(\sigma_0 + \eta T \max_{k\in[K]} \rho_k + \eta T \sigma_\xi d^{-1/2}\right)^{-\frac{q-1}{q}}\right\}\right),$$

for some $T \ge \widetilde{\Omega}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$*. For all* $t \le o(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2})$ *and* $i \in [n]$*,* $\max_{c\in[C]} y^{(i)} \left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle \le \widetilde{O}(\sigma_0\sigma_\xi)$.

Finally, Lemma 18 bounds $\langle \mathbf{w}_c(t), \boldsymbol{\xi} \rangle$ for some noise patch $\boldsymbol{\xi}$ from the testing set. Lemma 18 is useful in proving the test accuracy.

Lemma 18 (Bound on $\langle \mathbf{w}_c(t), \boldsymbol{\xi} \rangle$ for $\boldsymbol{\xi}$ from the testing set)**.** *Let* $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_\xi^2 I_d)$ *be a random noise vector independent of the dataset. Suppose* $\mathcal{G}_{\text{init}}$ *holds,* $C = \Theta(\log d)$*,* $n \le o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^q d^{1/2}, \sigma_0^{q-1}\sigma_\xi^{q-1} d^{1/2}\right\}\right)$*,* $K \le o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^{-1} d^{1/2}, \sigma_0^{q-1} d^{1/2}\right\}\right)$*, and*

$$\alpha \le o\left(P^{-\frac{1}{q}} \min\left\{1, \sigma_\xi^{\frac{1}{q}} d^{-\frac{1}{2q}} \left(\sigma_0 + \eta T \max_{k\in[K]} \rho_k + \eta T \sigma_\xi d^{-1/2}\right)^{-\frac{q-1}{q}}, \left(\sigma_0 + \eta T \max_{k\in[K]} \rho_k + \eta T \sigma_\xi d^{-1/2}\right)^{-1}\right\}\right),$$

for some $T = \widetilde{\Theta}\left(\max\left\{n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}, K\eta^{-1}\sigma_0^{-q+2}\right\}\right)$*. With probability at least* $1 - \frac{nK}{\text{poly}d}$*, for all* $c \in [C]$ *and* $0 \le t \le T$*,*

$$|\langle \mathbf{w}_c(t), \boldsymbol{\xi} \rangle - \langle \mathbf{w}_c(0), \boldsymbol{\xi} \rangle| \le o(\sigma_0\sigma_\xi).$$

F.2. Proof of main results from Lemmas in Appendix F.1

We first derive some implications of Assumption 2 that we use as conditions in the lemmas in F.1.

1. $nK \leq o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^q d^{1/2}, \sigma_0^{q-1}\sigma_\xi^{q-1}d^{1/2}\right\}\right)$ follows from $nK \leq o(\sigma_0^{q-1}\sigma_\xi^{q-1}d^{1/2})$ and $\sigma_\xi \geq \omega(1)$.
2. $K \leq o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^{-1}d^{1/2}, \sigma_0^{q-1}d^{1/2}\right\}\right)$ follows from $nK \leq o(\sigma_0^{q-1}\sigma_\xi^{q-1}d^{1/2})$, $\sigma_\xi \geq \omega(1)$ and $n \geq \omega(\sigma_\xi^q)$.
3. $\sigma_\xi d^{-1/2} \leq o(1)$ follows from $nK \leq o\left(\sigma_0^{q-1}\sigma_\xi^{q-1}d^{1/2}\right)$, $\sigma_0\sigma_\xi \leq o(1)$ and $n \geq \omega(\sigma_\xi^q)$.
4. $\sigma_\xi^q K \leq o(d^{1/2})$ follows from $nK \leq o(\sigma_0^{q-1}\sigma_\xi^{q-1}d^{1/2})$, $\sigma_\xi\sigma_0 \leq o(1)$ and $o(n) \geq \sigma_\xi^q \geq \omega(1)$.
5. $\alpha \leq o\left(P^{-\frac{1}{q}}\sigma_\xi \min\left\{d^{-1/2}, \sigma_0\right\}\left(\sigma_0 + \eta T \max_{k\in[K]}\rho_k + \eta T\sigma_\xi d^{-1/2}\right)^{-1}\right)$ follows from $\sigma_\xi d^{-1/2} \leq o(1)$, $\sigma_0 \leq o(1)$ and $\eta T \geq \omega(1)$.

Now, using Lemma 11 - 18, we prove the main theorems.

Theorem 3 (Training without data augmentation)**.** *Suppose that Assumption 2 holds. Let $\overline{T}$ be the first time step such that $\mathbf{w}(\overline{T})$ can classify all $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}_{train}$ with constant margin, i.e., ,*

$$y^{(i)}F(\mathbf{w}(\overline{T}), \mathbf{x}^{(i)}) \geq \widetilde{\Omega}(1), \quad \textit{for all } (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}_{train}.$$

For hidden channel number $C = \Theta(\log d)$, and small step size η, with probability at least $1 - O(\frac{n^2K}{\text{poly}(d)})$, $\overline{T} = \widetilde{\Theta}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$. Moreover, at time step $\overline{T}$, views $\mathbf{v}_2, \ldots, \mathbf{v}_K$ have never been learned, so that $\forall_{0\leq t\leq\overline{T}}$,

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}}\left[yF(\mathbf{w}(t), \mathbf{x}) < 0\right] \geq \left(\frac{1}{2} - O\left(\frac{1}{\sqrt{C}}\right)\right)\sum_{k=2}^{K}\rho_k.$$

Proof. By Lemma 1, with probability at least $1 - O\left(\frac{n^2K\log d}{\text{poly}d}\right)$, $\mathcal{G}_{\text{init}}$ holds. We first show that all $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}_{\text{train}}$ can be classified correctly with constant margin at some $T = \widetilde{\Theta}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$. We first consider the samples $i \in [n]$ such that $k_i^* = 1$. If Assumption 2 holds, $\omega(\sigma_\xi^q) \leq n$, so $\eta^{-1}\rho_1^{-1}\sigma_0^{-q+2} \leq o\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$. By Lemma 17, $\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle \leq \widetilde{O}(\sigma_0\sigma_\xi)$ for all $t \leq \widetilde{O}\left(\eta^{-1}\rho_1\sigma_0^{-q+2}\right)$. Then, by Lemma 14, there exists some $t^* \leq \widetilde{O}\left(\eta^{-1}\rho_1^{-1}\sigma_0^{-q+2}\right)$ such that $\max_{c\in[C]}\langle \mathbf{w}_c(t^*), \mathbf{v}_1\rangle = \Theta\left(C^{-1/q}\right)$. Moreover, by Lemma 13, at any time step $t^* \leq t' \leq \widetilde{O}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$, the feature $\mathbf{v}_1$ satisfies,

$$\max_{c\in[C]}\langle \mathbf{w}_c(t'), \mathbf{v}_1\rangle \geq \max_{c\in[C]}\langle \mathbf{w}_c(t^*), \mathbf{v}_1\rangle - o\left(\sigma_0\right) \geq \Omega\left(C^{-1/q}\right).$$

We can further show for all $c \in [C]$ and $t' \leq \widetilde{O}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$, $\langle \mathbf{w}_c(t'), \mathbf{v}_1\rangle$ and $y^{(i)}\left\langle \mathbf{w}_c(t'), \boldsymbol{\xi}^{(i)}\right\rangle$ are lower bounded. By Lemma 13, when $\mathcal{G}_{\text{init}}$ holds,

$$\langle \mathbf{w}_c(t'), \mathbf{v}_1\rangle \geq \langle \mathbf{w}_c(0), \mathbf{v}_1\rangle - o\left(\sigma_0\right) \geq -\widetilde{O}(\sigma_0),$$

and for all $i \in [n]$,

$$y^{(i)}\left\langle \mathbf{w}_c(t'), \boldsymbol{\xi}^{(i)}\right\rangle \geq y^{(i)}\left\langle \mathbf{w}_c(0), \boldsymbol{\xi}^{(i)}\right\rangle - o\left(\sigma_0\sigma_\xi\right) \geq -\widetilde{O}(\sigma_0\sigma_\xi).$$

Then, there exists some $T = \widetilde{\Theta}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$ such that for i with $k_i^* = 1$,

$$y^{(i)}F(\mathbf{w}(T), \mathbf{x}^{(i)}) = y^{(i)}\sum_{c\in[C]}\sum_{p\in[P]}\psi\left(\left\langle \mathbf{w}_c(T), \mathbf{x}_p^{(i)}\right\rangle\right)$$

$$
\begin{aligned}
&= y^{(i)} \sum_{c\in[C]} \psi\left(\left\langle \mathbf{w}_c(T), y^{(i)}\mathbf{v}_{k_i^*}\right\rangle\right) + y^{(i)} \sum_{c\in[C]} \sum_{k\in[K]} \sum_{p\in\mathcal{P}_{bp,k}^{(i)}} \psi\left(\left\langle \mathbf{w}_c(T), -\alpha_{p,i} y^{(i)}\mathbf{v}_k\right\rangle\right) \\
&\quad + y^{(i)} \sum_{c\in[C]} \psi\left(\left\langle \mathbf{w}_c(T), \boldsymbol{\xi}^{(i)}\right\rangle\right) \qquad (12) \\
&\geq \Omega\left(\frac{1}{C}\right) - C\widetilde{O}(\sigma_0^q) - CP\alpha^q\widetilde{O}\left(\left(\sigma_0 + \eta T\left(\max_{k\in[K]}\rho_k + \sigma_\xi d^{-1/2}\right)\right)^q\right) - C\widetilde{O}(\sigma_0^q\sigma_\xi^q) \\
&\geq \widetilde{\Omega}\left(1\right).
\end{aligned}
$$

The third step follows from $\max_{c\in[C]}\langle \mathbf{w}_c(T), \mathbf{v}_1\rangle \geq \Omega\left(C^{-1/q}\right)$, $\min_{c\in[C]}\langle \mathbf{w}_c(T), \mathbf{v}_1\rangle \geq -\widetilde{O}(\sigma_0)$, $\min_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(T), \boldsymbol{\xi}^{(i)}\right\rangle \geq -\widetilde{O}(\sigma_0\sigma_\xi)$ and Lemma 11. The last step follows from the the upper bound assumption on α, $\sigma_0 \leq o(1)$ and $\sigma_0\sigma_\xi \leq o(1)$.

We next show that the training accuracy is perfect for all $i \in [n]$ such that $k_i^* \neq 1$. By Lemma 16 and Assumption 2 that $\rho_k \leq o\left(n^{-1}\sigma_\xi^q\right)$ and $n \leq o(\sigma_\xi^{q-1}\sigma_0^{q-1}d^{1/2})$, we have $\frac{\sigma_0}{\eta\rho_k\sigma_0^{q-1}+\eta\sigma_\xi d^{-1/2}} \geq \omega\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$, and therefore $\max_{c\in[C]}\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle \leq \widetilde{O}\left(\sigma_0\right)$ for all $0 \leq t \leq \widetilde{O}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$ and $k \neq 1$. Then, for any $i \in [n]$ such that $k_i^* \neq 1$, by Lemma 15, there exists some time step $t^{(i)}$ such that $\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t^{(i)}), \boldsymbol{\xi}^{(i)}\right\rangle \geq \Omega\left(C^{-1/q}\right)$. Moreover, by Lemma 13, for all $t^{(i)} \leq t' \leq \widetilde{O}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$, $\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t'), \boldsymbol{\xi}^{(i)}\right\rangle \geq \Omega\left(C^{-1/q}\right)$.

Then, there exists some $T = \widetilde{\Theta}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$ such that for all $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}_{\text{train}}$ such that $k_i^* \neq 1$,

$$
\begin{aligned}
y^{(i)}F(\mathbf{w}(T), \mathbf{x}^{(i)}) &\geq \Omega\left(\frac{1}{C}\right) - C\widetilde{O}(\sigma_0^q) - CP\alpha^q\widetilde{O}\left(\left(\sigma_0 + \eta T\left(\max_{k\in[K]}\rho_k + \sigma_\xi d^{-1/2}\right)\right)^q\right) - C\widetilde{O}(\sigma_0^q\sigma_\xi^q) \\
&\geq \Omega\left(\frac{1}{C}\right).
\end{aligned}
$$

The first step follows from (12), and Lemma 11. The second step follows from the upper bound assumption on α, $\sigma_0 \leq o(1)$ and $\sigma_0\sigma_\xi \leq o(1)$.

Thus, at some $T = \widetilde{\Theta}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$, for all $i \in [n]$, we have $y^{(i)}F(\mathbf{w}(t), \mathbf{x}^{(i)}) \geq \Omega\left(\frac{1}{C}\right) \geq \widetilde{\Omega}\left(1\right)$.

Next, we show that the margin is $o(1)$ at $t \leq o\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$ for any $(\mathbf{x}^{(i)}, y^{(i)})$ such that $k_i^* \neq 1$. Since $t \leq o\left(\frac{\sigma_0}{\eta\rho_k\sigma_0^{q-1}+\eta\sigma_\xi d^{-1/2}}\right)$, by Lemma 16, $\max_{c\in[C]}\langle \mathbf{w}_c(t), \mathbf{v}_{k_i^*}\rangle \leq \widetilde{O}\left(\sigma_0\right)$. Since $t \leq o\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$, by Lemma 17, $y^{(i)}\left\langle \mathbf{w}_c(T), \boldsymbol{\xi}^{(i)}\right\rangle \leq \widetilde{O}(\sigma_0\sigma_\xi)$. Then,

$$
\begin{aligned}
y^{(i)}F(\mathbf{w}(t), \mathbf{x}^{(i)}) &\leq C\widetilde{O}(\sigma_0^q) + CP\alpha^q\widetilde{O}\left(\left(\sigma_0 + \eta T\left(\max_{k\in[K]}\rho_k + \sigma_\xi d^{-1/2}\right)\right)^q\right) + C\widetilde{O}(\sigma_0^q\sigma_\xi^q) \\
&\leq o(1). \qquad (13)
\end{aligned}
$$

The first step follows from (12). The second step follows from the upper bound assumption on α, $\sigma_0 \leq o(1)$ and $\sigma_0\sigma_\xi \leq o(1)$. Thus, we have show that $\overline{T} = \widetilde{\Theta}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$.

Finally, we show that the testing accuracy is bad on the testing dataset. For any $(\mathbf{x}, y) \sim \mathcal{D}$ with the main feature $\mathbf{v}_{k^*}$ such that $k^* \neq 1$ and dominant noise $\boldsymbol{\xi}$, since $\max_{c\in[C]}|\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle| \leq \widetilde{O}\left(\sigma_0\right)$ for any $t \leq \overline{T}$, following (12),

$$
\begin{aligned}
yF(\mathbf{w}(t), \mathbf{x}) &\leq C\widetilde{O}(\sigma_0^q) + CP\alpha^q\widetilde{O}\left(\left(\sigma_0 + \eta\overline{T}\left(\max_{k\in[K]}\rho_k + \sigma_\xi d^{-1/2}\right)\right)^q\right) + y\sum_{c\in[C]}\psi\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}\rangle\right) \\
&\leq C\widetilde{O}(\sigma_0^q) + C\widetilde{O}\left(\sigma_\xi^q\sigma_0^q\right)
\end{aligned}
$$

$$+ y \sum_{c\in[C]} \psi\left(\langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle\right) + \left| y \sum_{c\in[C]} \psi\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}\rangle\right) - y \sum_{c\in[C]} \psi\left(\langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle\right)\right|$$

$$\leq Co(\sigma_\xi^q \sigma_0^q) + y \sum_{c\in[C]} \psi\left(\langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle\right) + \left| y \sum_{c\in[C]} \psi\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}\rangle\right) - y \sum_{c\in[C]} \psi\left(\langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle\right)\right|.$$

The second step uses the upper bound on α. The last step follows the assumption $\sigma_\xi \geq \omega(1)$. For any $c \in [C]$, by Lemma 4, with probability at least $1 - \frac{1}{\text{poly}d}$, $|\langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle| \leq \widetilde{O}(\sigma_0\sigma_\xi)$. Then, by Lemma 18, $|\langle \mathbf{w}_c(t), \boldsymbol{\xi}\rangle - \langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle| \leq o(\sigma_0\sigma_\xi)$ with probability at least $1 - \frac{nK}{\text{poly}d}$ and therefore $|\langle \mathbf{w}_c(t), \boldsymbol{\xi}\rangle| \leq \widetilde{O}(\sigma_0\sigma_\xi)$ and

$$\left| y \sum_{c\in[C]} \psi\left(\langle \mathbf{w}_c(t), \boldsymbol{\xi}\rangle\right) - y \sum_{c\in[C]} \psi\left(\langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle\right)\right| \leq \sum_{c\in[C]} q\widetilde{O}(\sigma_0^{q-1}\sigma_\xi^{q-1}) \left|\langle \mathbf{w}_c(t), \boldsymbol{\xi}\rangle - \langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle\right|$$
$$\leq Co(\sigma_\xi^q \sigma_0^q)$$

For $t = 0$, $\langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle \sim \mathcal{N}(0, \sigma_0^2 \|\xi\|^2)$ and $\{\langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle : c \in [C]\}$ are independent. By Lemma 3, $\|\xi\|^2 = \Theta(\sigma_\xi^2)$. Then, by Lemma 9, with probability at least $\frac{1}{2} - O(\frac{1}{\sqrt{C}})$,

$$yF(\mathbf{w}(t), \mathbf{x}) \leq Co(\sigma_\xi^q \sigma_0^q) + y \sum_{c\in[C]} \psi\left(\langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle\right) < 0.$$

□

Theorem 4 (Training with data augmentation)**.** *Suppose assumption 2 holds. Let $\overline{T}_{aug}$ be the first time step such that $\mathbf{w}(\overline{T}_{aug})$ can classify all $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}_{train}^{(aug)}$ with constant margin, i.e.,*

$$y^{(i)} F(\mathbf{w}(\overline{T}_{aug}), \mathbf{x}^{(i)}) \geq \widetilde{\Omega}(1), \quad \textit{for all } (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}_{train}^{(aug)}.$$

For hidden channels number $C = \Theta(\log d)$, and small step size η, with probability at least $1 - O(\frac{n^2K^3}{\text{poly}(d)})$, $\overline{T}_{aug} = \widetilde{\Theta}\left(K\eta^{-1}\sigma_0^{-q+2}\right)$, and at $\overline{T}_{aug}$,

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}}\left[yF(\mathbf{w}(\overline{T}_{aug}), \mathbf{x}) < 0\right] \leq \frac{nK}{\text{poly}(d)}.$$

Proof. By Lemma 1a, $\mathcal{G}_{\text{init}}$ holds with probability at least $1 - O\left(\frac{n^2K^3 \log d}{\text{poly}d}\right)$.

We first show that $\overline{T}_{\text{aug}} = \widetilde{O}\left(K\eta^{-1}\sigma_0^{-q+2}\right)$. For the augmented dataset, we have $\rho_k^{(\text{aug})} = \frac{1}{K}$ for all $k \in [K]$ and the size of the dataset is Kn. For any $k \in [K]$, if Assumption 2 holds, $\omega(\sigma_\xi^q) \leq n$, so $\eta^{-1}\rho_k^{(\text{aug})-1}\sigma_0^{-q+2} \leq o\left(Kn\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$. Then, for any $i \in [Kn]$ with $k_i^* = k$, by Lemma 17 $\max_{c\in[C]} y^{(i)} \left\langle \mathbf{w}_c(T), \boldsymbol{\xi}^{(i)}\right\rangle \leq \widetilde{O}\left(\sigma_0\sigma_\xi\right)$ for all $0 \leq t \leq \widetilde{O}\left(K\eta^{-1}\sigma_0^{-q+2}\right)$. Then, under the assumption $\sigma_\xi^q K \leq o\left(d^{1/2}\right)$, by Lemma 14, there exists some $t_k = \widetilde{\Theta}\left(\frac{1}{\eta\rho_k^{(\text{aug})}\sigma_0^{q-2}}\right)$ such that $\max_{c\in[C]} \langle \mathbf{w}_c(t_k), \mathbf{v}_k\rangle \geq \Omega\left(C^{-1/q}\right)$. By Lemma 13, for any $t_k \leq t' \leq \widetilde{\Theta}\left(K\eta^{-1}\sigma_0^{-q+2}\right)$, $\max_{c\in[C]} \langle \mathbf{w}_c(t'), \mathbf{v}_k\rangle \geq \Omega\left(C^{-1/q}\right)$. Then, there exists some $T = \widetilde{\Theta}\left(K\eta^{-1}\sigma_0^{-q+2}\right)$ such that for all $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}_{\text{train}}^{(\text{aug})}$,

$$\begin{aligned} y^{(i)} F(\mathbf{w}(T), \mathbf{x}^{(i)}) &= y^{(i)} \sum_{c\in[C]} \sum_{p\in[P]} \psi\left(\left\langle \mathbf{w}_c(T), \mathbf{x}_p^{(i)}\right\rangle\right) \\ &= y^{(i)} \sum_{c\in[C]} \psi\left(\left\langle \mathbf{w}_c(T), y^{(i)}\mathbf{v}_{k_i^*}\right\rangle\right) + y^{(i)} \sum_{c\in[C]} \sum_{k\in[K]} \sum_{p\in\mathcal{P}_{bp,k}^{(i)}} \psi\left(\left\langle \mathbf{w}_c(T), -\alpha_{p,i}y^{(i)}\mathbf{v}_k\right\rangle\right) \\ &\quad + y^{(i)} \sum_{c\in[C]} \psi\left(\left\langle \mathbf{w}_c(T), \boldsymbol{\xi}^{(i)}\right\rangle\right) \end{aligned} \tag{14}$$

$$\geq \Omega\left(\frac{1}{C}\right) - C\widetilde{O}(\sigma_0^q) - CP\alpha^q\widetilde{O}\left(\left(\sigma_0 + \eta T\left(\max_{k\in[K]}\rho_k + \sigma_\xi d^{-1/2}\right)\right)^q\right) - C\widetilde{O}(\sigma_0^q\sigma_\xi^q)$$
$$\geq \widetilde{\Omega}(1).$$

The third step follows from $\max_{c\in[C]}\langle \mathbf{w}_c(T), \mathbf{v}_k\rangle \geq \Omega\left(C^{-1/q}\right)$, Lemma 11 and Lemma 13. The last step follows from the upper bound assumption on α, $\sigma_0 \leq o(1)$ and $\sigma_0\sigma_\xi \leq o(1)$.

Next, when $t = o\left(\frac{1}{\eta\rho_k^{(\text{aug})}\sigma_0^{q-2}}\right)$, by Lemma 16 , $\left\langle \mathbf{w}_c(t), \mathbf{v}_{k_i^*}\right\rangle \leq \widetilde{O}(\sigma_0)$. By Lemma 17, $y^{(i)}\left\langle \mathbf{w}_c(T), \boldsymbol{\xi}^{(i)}\right\rangle \leq \widetilde{O}(\sigma_0\sigma_\xi)$. Then,

$$y^{(i)}F(\mathbf{w}(t), \mathbf{x}^{(i)}) \leq C\widetilde{O}(\sigma_0^q) + CP\alpha^q\widetilde{O}\left(\left(\sigma_0 + \eta T\left(\max_{k\in[K]}\rho_k + \sigma_\xi d^{-1/2}\right)\right)^q\right) + C\widetilde{O}(\sigma_0^q\sigma_\xi^q)$$
$$\leq o(1).$$

The second step follows from the upper bound assumption on α, $\sigma_0 \leq o(1)$ and $\sigma_0\sigma_\xi \leq o(1)$. Thus, we have shown that $T^{(\text{aug})} = \widetilde{\Theta}\left(K\eta^{-1}\sigma_0^{-q+2}\right)$.

Finally, we show that the testing accuracy is perfect at $T^{(\text{aug})} = \widetilde{\Theta}\left(K\eta^{-1}\sigma_0^{-q+2}\right)$. For any sample $(\mathbf{x}, y)$ in the testing set with dominant noise $\boldsymbol{\xi}$, if $\mathcal{G}_{\text{init}}$ hold, by (14),

$$yF(\mathbf{w}(T^{(\text{aug})}), \mathbf{x}) \geq \Omega\left(\frac{1}{C}\right) - C\widetilde{O}(\sigma_0^q) - CP\alpha^q\widetilde{O}\left(\left(\sigma_0 + \eta T^{(\text{aug})}\left(\max_{k\in[K]}\rho_k + \sigma_\xi d^{-1/2}\right)\right)^q\right)$$
$$+ y\sum_{c\in[C]}\psi\left(\left\langle \mathbf{w}_c(T^{(\text{aug})}), \boldsymbol{\xi}\right\rangle\right)$$
$$\geq \Omega\left(\frac{1}{C}\right) + y\sum_{c\in[C]}\psi\left(\langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle\right) - \left|y\sum_{c\in[C]}\psi\left(\left\langle \mathbf{w}_c(T^{(\text{aug})}), \boldsymbol{\xi}\right\rangle\right) - y\sum_{c\in[C]}\psi\left(\langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle\right)\right|.$$

For any $c \in [C]$, by Lemma 4, with probability at least $1 - \frac{1}{\text{poly}d}$, $|\langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle| \leq \widetilde{O}(\sigma_0\sigma_\xi)$. Then, by Lemma 18, $\left|\left\langle \mathbf{w}_c(T^{(\text{aug})}), \boldsymbol{\xi}\right\rangle - \langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle\right| \leq o(\sigma_0\sigma_\xi)$ with probability at least $1 - \frac{nK}{\text{poly}d}$ and therefore $\left|\left\langle \mathbf{w}_c(T^{(\text{aug})}), \boldsymbol{\xi}\right\rangle\right| \leq \widetilde{O}(\sigma_0\sigma_\xi)$ and

$$\left|y\sum_{c\in[C]}\psi\left(\left\langle \mathbf{w}_c(T^{(\text{aug})}), \boldsymbol{\xi}\right\rangle\right) - y\sum_{c\in[C]}\psi\left(\langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle\right)\right| \leq \sum_{c\in[C]} q\widetilde{O}(\sigma_0^{q-1}\sigma_\xi^{q-1})\left|\left\langle \mathbf{w}_c(T^{(\text{aug})}), \boldsymbol{\xi}\right\rangle - \langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle\right|$$
$$\leq Co(\sigma_\xi^q\sigma_0^q).$$

Thus, with probability at least $1 - \frac{nK}{\text{poly}d}$, $yF(\mathbf{w}(T^{(\text{aug})}), \mathbf{x}) \geq \widetilde{\Omega}(1)$. □

G. Deferred Proof of Lemmas in Appendix F

In this section, we present the proof of lemmas necessary for proving our main result.

Lemma 11 (Rough upper and lower bound on $\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle$)**.** *Suppose $\mathcal{G}_{\text{init}}$ holds and*

$$\alpha \leq o\left(\sigma_\xi^{\frac{1}{q}} d^{-\frac{1}{2q}} P^{-\frac{1}{q}}\left(\sigma_0 + \eta T\rho_k + \eta T\sigma_\xi d^{-1/2}\right)^{-\frac{q-1}{q}}\right).$$

For all $0 \leq t' \leq t \leq T$ and $k \in [K]$, we have

$$\max_{c\in[C]}\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle \leq \max_{c\in[C]}\langle \mathbf{w}_c(t'), \mathbf{v}_k\rangle + \eta(t-t')\widetilde{O}\left(\rho_k + \sigma_\xi d^{-1/2}\right)$$
$$\leq \widetilde{O}\left(\sigma_0 + \eta T\left(\rho_k + \sigma_\xi d^{-1/2}\right)\right),$$

and

$$\min_{c\in[C]} \langle \mathbf{w}_c(t), \mathbf{v}_k\rangle \geq \min_{c\in[C]} \langle \mathbf{w}_c(t'), \mathbf{v}_k\rangle - \eta(t-t')\widetilde{O}\left(\sigma_\xi d^{-1/2}\right)$$
$$\geq -\widetilde{O}\left(\sigma_0 + \eta T\sigma_\xi d^{-1/2}\right).$$

Proof. For any $k \in [K]$, $c \in [C]$ and $0 \leq t < T$,

$$\begin{aligned}
&\langle \mathbf{w}_c(t+1), \mathbf{v}_k\rangle \\
=&\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle + \frac{\eta}{n}\sum_{i:k_i^*=k}\left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}}\psi'\left(\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle\right)\|\mathbf{v}_k\|_2^2\right) \\
&-\frac{\eta}{n}\sum_{i=1}^n\left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}}\sum_{p\in\mathcal{P}_{bp,k}^{(i)}}\alpha_{p,i}\psi'\left(\langle \mathbf{w}_c(t), \alpha_{p,i}\mathbf{v}_k\rangle\right)\|\mathbf{v}_k\|_2^2\right) \\
&+\frac{\eta}{n}\sum_{i=1}^n\left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}}\psi'\left(\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle\right)y^{(i)}\left\langle \boldsymbol{\xi}^{(i)}, \mathbf{v}_k\right\rangle\right) \qquad (15)
\end{aligned}$$

We bound each term separately. Since $\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}} \leq 1$ for all $i \in [n]$, $\|\mathbf{v}_k\|_2^2 = 1$, and $\psi'\left(\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle\right) \leq 1$ for all $k \in [K]$, we have

$$\frac{\eta}{n}\sum_{i:k_i^*=k}\left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}}\psi'\left(\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle\right)\|\mathbf{v}_k\|_2^2\right) \leq O(\eta\rho_k).$$

The feature noise term

$$-\frac{\eta}{n}\sum_{i=1}^n\left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}}\sum_{p\in\mathcal{P}_{bp,k}^{(i)}}\alpha_{p,i}\psi'\left(\langle \mathbf{w}_c(t), \alpha_{p,i}\mathbf{v}_k\rangle\right)\|\mathbf{v}_k\|_2^2\right) \leq 0.$$

When $\mathcal{G}_{\text{init}}$ holds, $\left\langle \boldsymbol{\xi}^{(i)}, \mathbf{v}_k\right\rangle \leq \widetilde{O}(\sigma_\xi d^{-1/2})$ for all $i \in [n]$. Since $\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}} \leq 1$ and $\psi'\left(\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle\right) \leq 1$,

$$\frac{\eta}{n}\sum_{i=1}^n\left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}}\psi'\left(\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle\right)y^{(i)}\left\langle \boldsymbol{\xi}^{(i)}, \mathbf{v}_k\right\rangle\right) \leq \widetilde{O}\left(\eta\sigma_\xi d^{-1/2}\right).$$

Then, for all $0 \leq t < T$,

$$\langle \mathbf{w}_c(t+1), \mathbf{v}_k\rangle \leq \langle \mathbf{w}_c(t), \mathbf{v}_k\rangle + \eta\widetilde{O}\left(\rho_k + \sigma_\xi d^{-1/2}\right),$$

which implies for any $0 \leq t' \leq t \leq T$,

$$\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle \leq \langle \mathbf{w}_c(t'), \mathbf{v}_k\rangle + \eta(t-t')\widetilde{O}\left(\rho_k + \sigma_\xi d^{-1/2}\right).$$

Next, we lower bound $\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle$ using induction. When $\mathcal{G}_{\text{init}}$ holds, $\langle \mathbf{w}_c(0), \mathbf{v}_k\rangle \geq -\widetilde{O}\left(\sigma_0 + \eta T\sigma_\xi d^{-1/2}\right)$. Assume for all $0 \leq t' \leq t$,

$$\min_{c\in[C]} \langle \mathbf{w}_c(t), \mathbf{v}_k\rangle \geq \min_{c\in[C]} \langle \mathbf{w}_c(t'), \mathbf{v}_k\rangle - \eta(t-t')\widetilde{O}\left(\sigma_\xi d^{-1/2}\right)$$
$$\geq -\widetilde{O}\left(\sigma_0 + \eta T\sigma_\xi d^{-1/2}\right)$$

for induction. We have

$$\frac{\eta}{n}\sum_{i:k_i^*=k}\left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}}\psi'\left(\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle\right)\|\mathbf{v}_k\|_2^2\right) \geq 0.$$

We have shown that $\langle \mathbf{w}_c(t), \mathbf{v}_k \rangle \leq \widetilde{O}\left(\sigma_0 + \eta T \rho_k + \eta T \sigma_\xi d^{-1/2}\right)$ for all $c \in [C]$ and $k \in [K]$. By the induction hypothesis,

$$\begin{aligned}
&\frac{\eta}{n} \sum_{i=1}^{n} \left(\frac{1}{1 + e^{y^{(i)} F(\mathbf{w}(t), \mathbf{x}^{(i)})}} \sum_{p \in \mathcal{P}_{bp,k}^{(i)}} \alpha_{p,i} \psi'\left(\langle \mathbf{w}_c(t), \alpha_{p,i} \mathbf{v}_k \rangle\right) \|\mathbf{v}_k\|_2^2 \right) \\
&\leq \eta \alpha^q P \widetilde{O}\left(\left(\sigma_0 + \eta T \rho_k + \eta T \sigma_\xi d^{-1/2}\right)^{q-1} \right) \\
&\leq \widetilde{O}\left(\eta \sigma_\xi d^{-1/2}\right).
\end{aligned}$$

The last inequality follows from $\alpha \leq \widetilde{O}\left(\sigma_\xi^{\frac{1}{q}} d^{-\frac{1}{2q}} P^{-\frac{1}{q}} / \left(\sigma_0 + \eta T \rho_k + \eta T \sigma_\xi d^{-1/2}\right)^{\frac{q-1}{q}} \right)$. When $\mathcal{G}_{\text{init}}$ holds,

$$-\frac{\eta}{n} \sum_{i=1}^{n} \left(\frac{1}{1 + e^{y^{(i)} F(\mathbf{w}(t), \mathbf{x}^{(i)})}} \psi'\left(\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle \right) y^{(i)} \left\langle \boldsymbol{\xi}^{(i)}, \mathbf{v}_k \right\rangle \right) \leq \widetilde{O}\left(\eta \sigma_\xi d^{-1/2}\right).$$

Then, plugging into (15), for any $0 \leq t' \leq t+1 \leq T$,

$$-\langle \mathbf{w}_c(t+1), \mathbf{v}_k \rangle \leq -\langle \mathbf{w}_c(t'), \mathbf{v}_k \rangle + \eta(t+1-t') \widetilde{O}\left(\sigma_\xi d^{-1/2}\right).$$

Thus, we have completed the induction and therefore

$$\min_{c \in [C]} \langle \mathbf{w}_c(t), \mathbf{v}_k \rangle \geq \min_{c \in [C]} \langle \mathbf{w}_c(t'), \mathbf{v}_k \rangle - \eta(t-t') \widetilde{O}\left(\sigma_\xi d^{-1/2}\right).$$

Finally, for $t' = 0$, when $\mathcal{G}_{\text{init}}$ holds, $|\langle \mathbf{w}_c(0), \mathbf{v}_k \rangle| \leq \widetilde{O}(\sigma_0)$. □

Lemma 12 (Rough lower bound on $\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \rangle$)**.** *Suppose $\mathcal{G}_{\text{init}}$ holds and*

$$\alpha \leq \widetilde{O}\left(\min\left\{1, \sigma_\xi^{\frac{1}{q}} d^{-\frac{1}{2q}}\right\} P^{-1/q} \left(\sigma_0 + \eta T \left(\max_{k \in [K]} \rho_k + \sigma_\xi d^{-1/2} \right)\right)^{-(q-1)/q} \right).$$

For all $0 \leq t \leq t' \leq T$ and $i \in [n]$, we have

$$\min_{c \in [C]} y^{(i)} \left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle \geq \min_{c \in [C]} y^{(i)} \left\langle \mathbf{w}_c(t'), \boldsymbol{\xi}^{(i)} \right\rangle - \eta(t-t') \widetilde{O}\left(\sigma_\xi^2 d^{-1/2} + \sigma_\xi d^{-1/2}\right).$$

Proof. For any $c \in [C]$ and $i \in [n]$, we have

$$\begin{aligned}
&y^{(i)} \left\langle \mathbf{w}_c(t+1), \boldsymbol{\xi}^{(i)} \right\rangle \\
&= y^{(i)} \left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle + \frac{\eta}{n} \frac{1}{1 + e^{y^{(i)} F(\mathbf{w}(t), \mathbf{x}^{(i)})}} \psi'\left(\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle \right) \left\| \boldsymbol{\xi}^{(i)} \right\|_2^2 \\
&\quad + \frac{\eta}{n} \sum_{j: j \neq i} \left(\frac{y^{(i)} y^{(j)}}{1 + e^{y^{(j)} F(\mathbf{w}(t), \mathbf{x}^{(j)})}} \psi'\left(\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(j)} \right\rangle \right) \left\langle \boldsymbol{\xi}^{(j)}, \boldsymbol{\xi}^{(i)} \right\rangle \right) \qquad (16) \\
&\quad + \frac{\eta}{n} \sum_{j=1}^{n} \left(\frac{y^{(i)}}{1 + e^{y^{(j)} F(\mathbf{w}(t), \mathbf{x}^{(j)})}} \psi'\left(\left\langle \mathbf{w}_c(t), \mathbf{v}_{k_j^*} \right\rangle \right) \left\langle \mathbf{v}_{k_j^*}, \boldsymbol{\xi}^{(i)} \right\rangle \right) \qquad (17) \\
&\quad - \frac{\eta}{n} \sum_{j=1}^{n} \left(\frac{y^{(i)}}{1 + e^{y^{(j)} F(\mathbf{w}(t), \mathbf{x}^{(j)})}} \sum_{k \in [K]} \sum_{p \in \mathcal{P}_{bp,k}^{(j)}} \psi'\left(\langle \mathbf{w}_c(t), \alpha_{p,j} \mathbf{v}_k \rangle\right) \left\langle \alpha_{p,j} \mathbf{v}_k, \boldsymbol{\xi}^{(i)} \right\rangle \right) \qquad (18)
\end{aligned}$$

We have $\frac{\eta}{n} \left(\frac{1}{1 + e^{y^{(i)} F(\mathbf{w}(t), \mathbf{x}^{(i)})}} \psi'\left(\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle \right) \left\| \boldsymbol{\xi}^{(i)} \right\|_2^2 \right)$ positive for any $i \in [n]$. Since $\frac{1}{1 + e^{y^{(j)} F(\mathbf{w}(t), \mathbf{x}^{(j)})}} \leq 1$ and $\psi'\left(\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(j)} \right\rangle \right) \leq 1$ for all $j \in [n]$, if $\mathcal{G}_{\text{init}}$ holds,

$$(16) \geq -\eta\widetilde{O}\left(\sigma_\xi^2 d^{-1/2}\right), \quad (17) \geq -\eta\widetilde{O}\left(\sigma_\xi d^{-1/2}\right).$$

Also,

$$\begin{aligned}(18) &\geq -\eta\widetilde{O}\left(\alpha^q P\sigma_\xi d^{-1/2}\max_{k\in[K]}|\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle|^{q-1}\right)\\ &\geq -\eta\widetilde{O}\left(\alpha^q P\sigma_\xi d^{-1/2}\left(\sigma_0 + \eta T\left(\rho_k + \sigma_\xi d^{-1/2}\right)\right)^{q-1}\right)\\ &\geq -\eta\widetilde{O}\left(\sigma_\xi d^{-1/2}\right)\end{aligned}$$

The second inequality follows from Lemma 11. The third inequality follows from the upper bound on α. Then,

$$y^{(i)}\left\langle \mathbf{w}_c(t+1), \boldsymbol{\xi}^{(i)}\right\rangle - y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle \geq -\eta\widetilde{O}\left(\sigma_\xi^2 d^{-1/2} + \sigma_\xi d^{-1/2}\right),$$

which gives

$$\min_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle \geq \min_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t'), \boldsymbol{\xi}^{(i)}\right\rangle - \eta(t-t')\widetilde{O}\left(\sigma_\xi^2 d^{-1/2} + \sigma_\xi d^{-1/2}\right).$$

□

Lemma 13 (Lower bound on $\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle$ and $\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\rangle$)**.** *Suppose $\mathcal{G}_{\text{init}}$ holds,* $n \leq o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^q d^{1/2}, \sigma_0^{q-1}\sigma_\xi^{q-1} d^{1/2}\right\}\right)$, $K \leq o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^{-1} d^{1/2}, \sigma_0^{q-1} d^{1/2}\right\}\right)$, *and*

$$\alpha \leq \widetilde{O}\left(\min\left\{1, \sigma_\xi^{\frac{1}{q}} d^{-\frac{1}{2q}}\right\} P^{-1/q}\left(\sigma_0 + \eta T\left(\max_{k\in[K]}\rho_k + \sigma_\xi d^{-1/2}\right)\right)^{-(q-1)/q}\right).$$

for some $T = \widetilde{\Theta}\left(\max\left\{n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}, K\eta^{-1}\sigma_0^{-q+2}\right\}\right)$. *For all* $0 \leq t' \leq t \leq T$, *and* $c \in [C]$,

$$\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle \geq \langle \mathbf{w}_c(t'), \mathbf{v}_k\rangle - o(\sigma_0),$$

and for all $i \in [n]$,

$$y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle \geq y^{(i)}\left\langle \mathbf{w}_c(t'), \boldsymbol{\xi}^{(i)}\right\rangle - o\left(\sigma_0\sigma_\xi\right).$$

Proof. By Lemma 12, for any $(\mathbf{x}^{(i)}, y^{(i)})$,

$$\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle \geq \max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t'), \boldsymbol{\xi}^{(i)}\right\rangle - \eta(t-t')\widetilde{O}\left(\sigma_\xi^2 d^{-1/2} + \sigma_\xi d^{-1/2}\right).$$

Then, when $t - t' \leq \widetilde{O}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$ and $n \leq o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^q d^{1/2}, \sigma_0^{q-1}\sigma_\xi^{q-1} d^{1/2}\right\}\right)$, or when $t - t' \leq \widetilde{O}\left(K\eta^{-1}\sigma_0^{-q+2}\right)$ and $K \leq o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^{-1} d^{1/2}, \sigma_0^{q-1} d^{1/2}\right\}\right)$, $\eta(t-t')\widetilde{O}\left(\sigma_\xi^2 d^{-1/2} + \sigma_\xi d^{-1/2}\right) \leq o\left(\sigma_0\sigma_\xi\right)$.

By Lemma 11,

$$\max_{c\in[C]}\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle \geq \max_{c\in[C]}\langle \mathbf{w}_c(t'), \mathbf{v}_k\rangle - \eta(t-t')\widetilde{O}\left(\sigma_\xi d^{-1/2}\right).$$

Then, when $t - t' \leq \widetilde{O}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$ and $n \leq o\left(\sigma_0^{q-1}\sigma_\xi^{q-1} d^{1/2}\right)$, or when $t - t' \leq \widetilde{O}\left(K\eta^{-1}\sigma_0^{-q+2}\right)$ and $K \leq o(\sigma_0^{q-1}\sigma_\xi^{-1} d^{1/2})$,

$$\eta(t-t')\widetilde{O}\left(\sigma_\xi d^{-1/2}\right) \leq o\left(\sigma_0\right),$$

which completes the proof. □

Lemma 14 (Learning the main feature)**.** *Suppose $\mathcal{G}_{\text{init}}$ holds, $C = \Theta(\log d)$, $\sigma_0\sigma_\xi \le o(1)$, $\sigma_\xi^q d^{-1/2} \le o(\rho_k)$ and*

$$\alpha \le o\left(P^{-\frac{1}{q}} \min\left\{1, \sigma_\xi^{\frac{1}{q}} d^{-\frac{1}{2q}} \left(\sigma_0 + \eta T \max_{k\in[K]} \rho_k + \eta T \sigma_\xi d^{-1/2}\right)^{-\frac{q-1}{q}}, \left(\sigma_0 + \eta T \max_{k\in[K]} \rho_k + \eta T \sigma_\xi d^{-1/2}\right)^{-1}\right\}\right),$$

for some $T \ge \widetilde{\Omega}\left(\left(\eta\rho_k\sigma_0^{q-2}\right)^{-1}\right)$. For any $k \in [K]$ and $0 \le t \le T$, if

$$\max_{c\in[C]} \langle \mathbf{w}_c(t), \mathbf{v}_k\rangle \le O(C^{-1/q}), \quad \text{and} \quad \max_{i\in[n],c\in[C]} y^{(i)} \left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle \le \widetilde{O}(\sigma_0\sigma_\xi),$$

then

$$\max_{c\in[C]} \langle \mathbf{w}_c(t+1), \mathbf{v}_k\rangle = \max_{c\in[C]} \langle \mathbf{w}_c(t), \mathbf{v}_k\rangle + \Theta\left(\eta\rho_k\psi'\left(\max_{c\in[C]} \langle \mathbf{w}_c(t), \mathbf{v}_k\rangle\right)\right).$$

Moreover, if $\max_{i\in[n],c\in[C]} \left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle \le \widetilde{O}(\sigma_0\sigma_\xi)$ for all $t \le \widetilde{O}\left(\frac{1}{\eta\rho_k\sigma_0^{q-2}}\right)$, there exists $T' \le \widetilde{O}\left(\frac{1}{\eta\rho_k\sigma_0^{q-2}}\right)$ such that $\max_{c\in[C]} \langle \mathbf{w}_c(T'), \mathbf{v}_k\rangle \ge \Omega\left(C^{-1/q}\right)$.

Proof. By the upper bound on α and Lemma 11, for any $i \in [n]$ and $c \in [C]$,

$$\begin{aligned}
\sum_{k'\in[K]} \sum_{p\in\mathcal{P}_{bp,k'}^{(i)}} \psi\left(\left\langle \mathbf{w}_c(t), -y^{(i)}\alpha_{p,i}\mathbf{v}_{k'}\right\rangle\right) &\le \sum_{k'\in[K]} \sum_{p\in\mathcal{P}_{bp,k}^{(i)}} |\langle \mathbf{w}_c(t), \alpha_{p,i}\mathbf{v}_{k'}\rangle|^q \\
&\le \widetilde{O}\left(\alpha^q P\left(\sigma_0 + \eta T\left(\max_{k'} \rho_{k'} + \sigma_\xi d^{-1/2}\right)\right)^q\right) \\
&\le o(1).
\end{aligned}$$

Then, since $\max_{c\in[C]} \langle \mathbf{w}_c(t), \mathbf{v}_k\rangle \le O(C^{-1/q})$, and $\max_{c\in[C]} y\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle \le o(1)$ for all $i \in [n]$, we have for all i such that $k_i^* = k$, $y^{(i)}F(\mathbf{w}(t), \mathbf{x}^{(i)}) \le O(1)$ and $\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}} \ge \Omega(1)$.

Now, we compute the update $\langle \mathbf{w}_c(t+1), \mathbf{v}_k\rangle - \langle \mathbf{w}_c(t), \mathbf{v}_k\rangle$,

$$\begin{aligned}
&\langle \mathbf{w}_c(t+1), \mathbf{v}_k\rangle \\
=\ &\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle + \frac{\eta}{n} \sum_{i:k_i^*=k} \left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}} \psi'\left(\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle\right) \|\mathbf{v}_k\|_2^2\right) \\
&- \frac{\eta}{n} \sum_{i=1}^n \left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}} \sum_{p\in\mathcal{P}_{bp,k}^{(i)}} \alpha_{p,i}\psi'\left(\langle \mathbf{w}_c(t), \alpha_{p,i}\mathbf{v}_k\rangle\right) \|\mathbf{v}_k\|_2^2\right) \qquad (19) \\
&+ \frac{\eta}{n} \sum_{i=1}^n \left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}} \psi'\left(\left\langle \mathbf{w}_c^{(t)}, \boldsymbol{\xi}^{(i)}\right\rangle\right) y^{(i)} \left\langle \boldsymbol{\xi}^{(i)}, \mathbf{v}_k\right\rangle\right) \qquad (20)
\end{aligned}$$

Then, when $\mathcal{G}_{\text{init}}$ holds, since $\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}} \ge \Omega(1)$ for all $i \in [n]$ such that $k_i^* = k$,

$$\frac{\eta}{n} \sum_{i:k_i^*=k} \left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}} \psi'\left(\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle\right) \|\mathbf{v}_k\|_2^2\right) = \Theta\left(\eta\rho_k \left|\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle\right|^{q-1}\right).$$

We can bound the term (19) as

$$|(19)| \le \widetilde{O}\left(\frac{\eta}{N} \sum_{i=1}^n \sum_{p\in\mathcal{P}_{bp,k}^{(i)}} \alpha_{p,i}\psi'\left(\langle \mathbf{w}_c(t), \alpha_{p,i}\mathbf{v}_k\rangle\right)\right)$$

$$\leq \widetilde{O}\left(\eta\rho_k\alpha^q P\left|\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle\right|^{q-1}\right) \leq o(\eta\rho_k\left|\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle\right|^{q-1}).$$

For the term (20), if $\mathcal{G}_{\text{init}}$ holds,

$$\begin{aligned}
|(20)| &\leq \frac{\eta}{n}\sum_{i=1}^{n}\left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}}\psi'\left(\left\langle \mathbf{w}_c^{(t)}, \boldsymbol{\xi}^{(i)}\right\rangle\right)\left|\left\langle \boldsymbol{\xi}^{(i)}, \mathbf{v}_k\right\rangle\right|\right) \\
&\leq \widetilde{O}\left(\frac{\eta}{n}\sum_{i=1}^{n}\left|\left\langle \mathbf{w}_c^{(t)}, \boldsymbol{\xi}^{(i)}\right\rangle\right|^{q-1}\left\langle \boldsymbol{\xi}^{(i)}, \mathbf{v}_k\right\rangle\right) \\
&\leq \widetilde{O}\left(\eta\sigma_0^{q-1}\sigma_\xi^q d^{-1/2}\right).
\end{aligned}$$

For $t=0$, if $\mathcal{G}_{\text{init}}$ holds, $\max_{c\in[C]}\langle \mathbf{w}_c(0), \mathbf{v}_k\rangle \geq \widetilde{\Omega}(\sigma_0)$. Then, if $\sigma_0^{q-1}\sigma_\xi^q d^{-1/2} \leq o(\rho_k\sigma_0^{q-1})$, we have

$$\max_{c\in[C]}\langle \mathbf{w}_c(t+1), \mathbf{v}_k\rangle = \max_{c\in[C]}\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle + \Theta\left(\eta\rho_k\psi'\left(\max_{c\in[C]}\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle\right)\right), \tag{21}$$

which shows $\max_{c\in[C]}\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle$ is increasing. Then, (21) holds for all t.

Starting from some $\langle \mathbf{w}_c(t'), \mathbf{v}_k\rangle$, the number of iterations it takes to reach $\max_{c\in[C]}\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle \geq 2\max_{c\in[C]}\langle \mathbf{w}_c(t'), \mathbf{v}_k\rangle$ is at most $O\left(\frac{\max_{c\in[C]}\langle \mathbf{w}_c(t'), \mathbf{v}_k\rangle}{\eta\rho_k\left(\max_{c\in[C]}\langle \mathbf{w}_c(t'), \mathbf{v}_k\rangle\right)^{q-1}}\right)$. Then, starting from $\Theta(\sigma_0)$, it takes at most

$$\widetilde{O}\left(\sum_{i=0}^{\infty}\frac{2^i\sigma_0}{\eta\rho_k(2^i\sigma_0)^{q-1}}\right) \leq \tilde{O}\left(\frac{1}{\eta\rho_k\sigma_0^{q-2}}\right)$$

time steps to reach $\max_{c\in[C]}\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle \geq \Omega\left(C^{-1/q}\right)$. □

Lemma 15 (Overfitting the dominant noise)**.** *Suppose* $\mathcal{G}_{\text{init}}$ *holds,* $C = \Theta(\log d)$, $n \leq o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^q d^{1/2}, \sigma_0^{q-1}\sigma_\xi^{q-1} d^{1/2}\right\}\right)$ *and*

$$\alpha \leq o\left(P^{-\frac{1}{q}}\min\left\{1, \sigma_\xi^{\frac{1}{q}}d^{-\frac{1}{2q}}\left(\sigma_0 + \eta T\max_{k\in[K]}\rho_k + \eta T\sigma_\xi d^{-1/2}\right)^{-\frac{q-1}{q}}, \left(\sigma_0 + \eta T\max_{k\in[K]}\rho_k + \eta T\sigma_\xi d^{-1/2}\right)^{-1}\right\}\right),$$

for some $T \geq \widetilde{\Omega}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$.

Let $i\in[n]$ *be some sample such that for all* $0\leq t\leq T$, $\max_{c\in[C]}\left\langle \mathbf{w}_c(t), \mathbf{v}_{k_i^*}\right\rangle \leq O(C^{-1/q})$. *For any time step* $0\leq t\leq T$ *, if*

$$\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle \leq O(C^{-1/q}),$$

we have

$$\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t+1), \boldsymbol{\xi}^{(i)}\right\rangle = \max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle + \frac{\eta}{n}\widetilde{\Theta}\left(\sigma_\xi^2\psi'\left(\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle\right)\right).$$

Moreover, there exists times step $T' \leq \widetilde{O}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$ *such that* $\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(T'), \boldsymbol{\xi}^{(i)}\right\rangle \geq \Omega\left(C^{-1/q}\right)$.

Proof. By the upper bound on α and Lemma 11, for any $c\in[C]$,

$$\left|\sum_{k'\in[K]}\sum_{p\in\mathcal{P}_{bp,k'}^{(i)}}\psi\left(\langle \mathbf{w}_c(t), \alpha_{p,i}\mathbf{v}_{k'}\rangle\right)\right| \leq \sum_{k'\in[K]}\sum_{p\in\mathcal{P}_{bp,k}^{(i)}}\left|\langle \mathbf{w}_c(t), \alpha_{p,i}\mathbf{v}_{k'}\rangle\right|^q$$

$$\leq \widetilde{O}\left(\alpha^q P\left(\sigma_0 + \eta T\left(\max_{k'} \rho_{k'} + \sigma_\xi d^{-1/2}\right)\right)^q\right)$$
$$\leq o(1).$$

For i, when $\max_{c\in[C]} \left\langle \mathbf{w}_c(t), v_{k_i^*}\right\rangle \leq O(C^{-1/q})$, $\max_{c\in[C]} y^{(i)} \left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle \leq O(C^{-1/q})$ and

$$\left|\sum_{k\in[K]} \sum_{p\in\mathcal{P}_{bp,k}^{(i)}} \psi\left(\langle \mathbf{w}_c(t), -y\alpha_{p,i}\mathbf{v}_k\rangle\right)\right| \leq o(1),$$

we have $y^{(i)}F(\mathbf{w}(t), \mathbf{x}^{(i)}) \leq O(1)$ and therefore $\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}} \geq \Omega(1)$. Then,

$$\begin{aligned}
& y^{(i)}\left\langle \mathbf{w}_c(t+1), \boldsymbol{\xi}^{(i)}\right\rangle \\
&= y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle + \frac{\eta}{n}\left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}}\psi'\left(\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle\right)\left\|\boldsymbol{\xi}^{(i)}\right\|_2^2\right) \\
&\quad + \frac{\eta}{n}\sum_{j:j\neq i}\left(\frac{y^{(i)}y^{(j)}}{1+e^{y^{(j)}F(\mathbf{w}(t),\mathbf{x}^{(j)})}}\psi'\left(\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(j)}\right\rangle\right)\left\langle \boldsymbol{\xi}^{(j)}, \boldsymbol{\xi}^{(i)}\right\rangle\right) \qquad (22) \\
&\quad + \frac{\eta}{n}\sum_{j=1}^{n}\left(\frac{y^{(i)}}{1+e^{y^{(j)}F(\mathbf{w}_c(t),\mathbf{x}^{(j)})}}\psi'\left(\left\langle \mathbf{w}_c(t), \mathbf{v}_{k_j^*}\right\rangle\right)\left\langle \mathbf{v}_{k_j^*}, \boldsymbol{\xi}^{(i)}\right\rangle\right) \qquad (23) \\
&\quad - \frac{\eta}{n}\sum_{j=1}^{n}\left(\frac{y^{(i)}}{1+e^{y^{(j)}F(\mathbf{w}_c(t),\mathbf{x}^{(j)})}}\sum_{k\in[K]}\sum_{p\in\mathcal{P}_{bp,k}^{(j)}}\psi'\left(\langle \mathbf{w}_c(t), \alpha_{p,j}\mathbf{v}_k\rangle\right)\left\langle \alpha_{p,j}\mathbf{v}_k, \boldsymbol{\xi}^{(i)}\right\rangle\right). \qquad (24)
\end{aligned}$$

If $\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}} \geq \Omega(1)$, and $\widetilde{\Omega}\left(\sigma_0\sigma_\xi\right) \leq \max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle$,

$$\frac{\eta}{n}\left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}}\psi'\left(\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle\right)\left\|\boldsymbol{\xi}^{(i)}\right\|_2^2\right) \geq \widetilde{\Omega}\left(\frac{\eta}{n}\left(\sigma_0\sigma_\xi\right)^{q-1}\sigma_\xi^2\right).$$

When $\mathcal{G}_{\text{init}}$ holds, since $\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}} \leq 1$, $\psi'\left(\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(j)}\right\rangle\right) \leq 1$, and $\psi'\left(\left\langle \mathbf{w}_c(t), \mathbf{v}_{k_j^*}\right\rangle\right) \leq 1$, $|(22)| \leq \widetilde{O}\left(\eta\sigma_\xi^2 d^{-1/2}\right)$ and $|(23)| \leq \widetilde{O}\left(\eta\sigma_\xi d^{-1/2}\right)$.

By Lemma 11 and the upper bound on α,

$$\begin{aligned}
|(24)| &\leq \widetilde{O}\left(\eta\alpha^q P\max_{k\in[K]}\left|\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle\right|^{q-1}\sigma_\xi d^{-1/2}\right) \\
&\leq \widetilde{O}\left(\eta\alpha^q P\left(\sigma_0 + \eta T\left(\rho_k + \sigma_\xi\right)\right)^{q-1}\sigma_\xi d^{-1/2}\right) \\
&\leq \widetilde{O}\left(\eta\sigma_\xi d^{-1/2}\right).
\end{aligned}$$

When $\mathcal{G}_{\text{init}}$ holds, $\widetilde{\Omega}\left(\sigma_0\sigma_\xi\right) \leq y^{(i)}\max_{c\in[C]}\left\langle \mathbf{w}_c(0), \boldsymbol{\xi}^{(i)}\right\rangle$. Then, when $n \leq o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^q d^{1/2}, \sigma_0^{q-1}\sigma_\xi^{q-1}d^{1/2}\right\}\right)$, for $t = 0$,

$$\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t+1), \boldsymbol{\xi}^{(i)}\right\rangle = \max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle + \frac{\eta}{n}\widetilde{\Theta}\left(\sigma_\xi^2\psi'\left(\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle\right)\right), \qquad (25)$$

which shows $\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle$ is increasing. Then, (25) holds for all $0 \leq t \leq T$.

Starting from $\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t'), \boldsymbol{\xi}^{(i)}\right\rangle$, the number of iterations it takes to reach $\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle \geq 2\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t'), \boldsymbol{\xi}^{(i)}\right\rangle$ is at most $O\left(\frac{n\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t'), \boldsymbol{\xi}^{(i)}\right\rangle}{\eta\sigma_\xi^2\left(\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t'), \boldsymbol{\xi}^{(i)}\right\rangle\right)^{q-1}}\right)$. Then, starting from $\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(0), \boldsymbol{\xi}^{(i)}\right\rangle \geq \widetilde{\Omega}\left(\sigma_0\sigma_\xi\right)$, it takes at most

$$T' \leq \widetilde{O}\left(\sum_{i=0}^{\infty} \frac{n2^i\sigma_0\sigma_\xi}{\eta\sigma_\xi^2(2^i\sigma_0\sigma_\xi)^{q-1}}\right) \leq \tilde{O}\left(\frac{n}{\eta\sigma_\xi^2(\sigma_0\sigma_\xi)^{q-2}}\right)$$

time steps to reach $\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(T'), \boldsymbol{\xi}^{(i)}\right\rangle \geq \Omega(C^{-1/q})$. □

Lemma 16 (Upper bound on $\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle$)**.** *If $\mathcal{G}_{\text{init}}$ holds, for all $k\in[K]$ and $t \leq o\left(\frac{\sigma_0}{\eta\rho_k\sigma_0^{q-1}+\eta\sigma_\xi d^{-1/2}}\right)$,*

$$\max_{c\in[C]} \langle \mathbf{w}_c(t), \mathbf{v}_k\rangle \leq \widetilde{O}\left(\sigma_0\right).$$

Proof. For every $k\in[K]$,

$$\begin{aligned}
&\langle \mathbf{w}_c(t+1), \mathbf{v}_k\rangle \\
=&\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle + \frac{\eta}{n}\sum_{i:k_i^*=k}\left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}}\psi'\left(\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle\right)\|\mathbf{v}_k\|_2^2\right) \\
&-\frac{\eta}{n}\sum_{i=1}^{n}\left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}}\sum_{p\in\mathcal{P}_{bp,k}^{(i)}}\alpha_{p,i}\psi'\left(\langle \mathbf{w}_c(t), \alpha_{p,i}\mathbf{v}_k\rangle\right)\|\mathbf{v}_k\|_2^2\right) \qquad (26)\\
&+\frac{\eta}{n}\sum_{i=1}^{n}\left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}}\psi'\left(\left\langle \mathbf{w}_c^{(t)}, \boldsymbol{\xi}^{(i)}\right\rangle\right)y^{(i)}\left\langle \boldsymbol{\xi}^{(i)}, \mathbf{v}_k\right\rangle\right) \qquad (27)
\end{aligned}$$

Then, since $\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}} \leq 1$ and $\psi'\left(\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle\right) \leq O\left(|\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle|^{q-1}\right)$,

$$\frac{\eta}{n}\sum_{i:k_i^*=k}\left(\frac{1}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}}\psi'\left(\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle\right)\|\mathbf{v}_k\|_2^2\right) \leq O\left(\eta\rho_k\left|\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle\right|^{q-1}\right).$$

The second term (26)≤ 0. For (27), since $\left|\frac{y^{(i)}}{1+e^{y^{(i)}F(\mathbf{w}(t),\mathbf{x}^{(i)})}}\right| \leq 1$ and $\psi'\left(\left\langle \mathbf{w}_c^{(t)}, \boldsymbol{\xi}^{(i)}\right\rangle\right) \leq 1$, if $\mathcal{G}_{\text{init}}$ holds, (27) $\leq \widetilde{O}\left(\sigma_\xi d^{-1/2}\right)$.

Finally, if $\mathcal{G}_{\text{init}}$ holds, $\langle \mathbf{w}_c(0), \mathbf{v}_k\rangle \leq \widetilde{O}\left(\sigma_0\right)$, so it takes at least $t \geq \widetilde{\Omega}\left(\frac{\sigma_0}{\eta\rho_k\sigma_0^{q-1}+\eta\sigma_\xi d^{-1/2}}\right)$ time steps to reach $\langle \mathbf{w}_c(t), \mathbf{v}_k\rangle \geq 2\langle \mathbf{w}_c(0), \mathbf{v}_k\rangle$. □

Lemma 17 (Upper bound on $\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\rangle$)**.** *Suppose $\mathcal{G}_{\text{init}}$ holds, $n \leq o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^q d^{1/2}, \sigma_0^{q-1}\sigma_\xi^{q-1}d^{1/2}\right\}\right)$ and*

$$\alpha \leq o\left(P^{-\frac{1}{q}}\min\left\{1, \sigma_\xi^{\frac{1}{q}}d^{-\frac{1}{2q}}\left(\sigma_0 + \eta T\max_{k\in[K]}\rho_k + \eta T\sigma_\xi d^{-1/2}\right)^{-\frac{q-1}{q}}\right\}\right),$$

for some $T \geq \widetilde{\Omega}\left(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}\right)$. For all $t \leq o(n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2})$ and $i\in[n]$, $\max_{c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)}\right\rangle \leq \widetilde{O}(\sigma_0\sigma_\xi)$.

Proof. We prove using induction. At $t=0$, when $\mathcal{G}_{\text{init}}$ holds, $\max_{i\in[n],c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(0), \boldsymbol{\xi}^{(i)}\right\rangle \leq \widetilde{O}(\sigma_0\sigma_\xi)$. Assume $\max_{i\in[n],c\in[C]} y^{(i)}\left\langle \mathbf{w}_c(t'), \boldsymbol{\xi}^{(i)}\right\rangle \leq \widetilde{O}(\sigma_0\sigma_\xi)$ for any $0\leq t'\leq t$ for induction. For any $c\in[C]$,

$$y^{(i)}\left\langle \mathbf{w}_c(t+1), \boldsymbol{\xi}^{(i)}\right\rangle$$

$$= y^{(i)} \left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle + \frac{\eta}{n} \left(\frac{1}{1 + e^{y^{(i)} F(\mathbf{w}(t), \mathbf{x}^{(i)})}} \psi' \left(\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle \right) \left\| \boldsymbol{\xi}^{(i)} \right\|_2^2 \right)$$

$$+ \frac{\eta}{n} \sum_{j:j \neq i} \left(\frac{y^{(i)} y^{(j)}}{1 + e^{y^{(j)} F(\mathbf{w}(t), \mathbf{x}^{(j)})}} \psi' \left(\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(j)} \right\rangle \right) \left\langle \boldsymbol{\xi}^{(j)}, \boldsymbol{\xi}^{(i)} \right\rangle \right) \tag{28}$$

$$+ \frac{\eta}{n} \sum_{j=1}^{n} \left(\frac{y^{(i)}}{1 + e^{y^{(j)} F(\mathbf{w}_c(t), \mathbf{x}^{(j)})}} \psi' \left(\left\langle \mathbf{w}_c(t), \mathbf{v}_{k_j^*} \right\rangle \right) \left\langle \mathbf{v}_{k_j^*}, \boldsymbol{\xi}^{(i)} \right\rangle \right) \tag{29}$$

$$- \frac{\eta}{n} \sum_{j=1}^{n} \left(\frac{y^{(i)}}{1 + e^{y^{(j)} F(\mathbf{w}_c(t), \mathbf{x}^{(j)})}} \sum_{k \in [K]} \sum_{p \in \mathcal{P}_{bp,k}^{(j)}} \psi' \left(\langle \mathbf{w}_c(t), \alpha_{p,j} \mathbf{v}_k \rangle \right) \left\langle \alpha_{p,j} \mathbf{v}_k, \boldsymbol{\xi}^{(i)} \right\rangle \right). \tag{30}$$

Then, when $\mathcal{G}_{\text{init}}$ holds, by $\frac{1}{1+e^{y^{(i)} F(\mathbf{w}(t), \mathbf{x}^{(i)})}} \leq 1$ and $\psi'(\cdot) \leq 1$,

$$\begin{aligned}
& y^{(i)} \left\langle \mathbf{w}_c(t+1), \boldsymbol{\xi}^{(i)} \right\rangle \\
& \leq y^{(i)} \left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle + \eta \widetilde{O}(n^{-1} \sigma_0^{q-1} \sigma_\xi^{q+1} + \sigma_\xi^2 d^{-1/2} + \sigma_\xi d^{-1/2} + \alpha^q P \sigma_\xi d^{-1/2} \left(\sigma_0 + \eta T \left(\rho_k + \sigma_\xi d^{-1/2} \right) \right)^{q-1}) \\
& \leq y^{(i)} \left\langle \mathbf{w}_c(0), \boldsymbol{\xi}^{(i)} \right\rangle + \eta t \widetilde{O}(n^{-1} \sigma_0^{q-1} \sigma_\xi^{q+1} + \sigma_\xi^2 d^{-1/2} + \sigma_\xi d^{-1/2}).
\end{aligned}$$

The last step uses the upper bound on α and the induction hypothesis. Since $n \leq o\left(\min\left\{ \sigma_0^{q-1} \sigma_\xi^q d^{1/2}, \sigma_0^{q-1} \sigma_\xi^{q-1} d^{1/2} \right\} \right)$ and $t \leq o(n \eta^{-1} \sigma_\xi^{-q} \sigma_0^{-q+2})$, $\max_{i \in [n], c \in [C]} y^{(i)} \left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle \leq \widetilde{O}(\sigma_0 \sigma_\xi)$.

□

Lemma 18 (Bound on $\langle \mathbf{w}_c(t), \boldsymbol{\xi} \rangle$ for $\boldsymbol{\xi}$ from the testing set)**.** *Let $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_\xi^2 I_d)$ be a random noise vector independent of the dataset. Suppose $\mathcal{G}_{\text{init}}$ holds, $C = \Theta(\log d)$, $n \leq o\left(\min\left\{ \sigma_0^{q-1} \sigma_\xi^q d^{1/2}, \sigma_0^{q-1} \sigma_\xi^{q-1} d^{1/2} \right\} \right)$, $K \leq o\left(\min\left\{ \sigma_0^{q-1} \sigma_\xi^{-1} d^{1/2}, \sigma_0^{q-1} d^{1/2} \right\} \right)$, and*

$$\alpha \leq o\left(P^{-\frac{1}{q}} \min\left\{ 1, \sigma_\xi^{\frac{1}{q}} d^{-\frac{1}{2q}} \left(\sigma_0 + \eta T \max_{k \in [K]} \rho_k + \eta T \sigma_\xi d^{-1/2} \right)^{-\frac{q-1}{q}}, \left(\sigma_0 + \eta T \max_{k \in [K]} \rho_k + \eta T \sigma_\xi d^{-1/2} \right)^{-1} \right\} \right),$$

for some $T = \widetilde{\Theta}\left(\max\left\{ n \eta^{-1} \sigma_\xi^{-q} \sigma_0^{-q+2}, K \eta^{-1} \sigma_0^{-q+2} \right\} \right)$. With probability at least $1 - \frac{nK}{\text{poly}d}$, for all $c \in [C]$ and $0 \leq t \leq T$,

$$|\langle \mathbf{w}_c(t), \boldsymbol{\xi} \rangle - \langle \mathbf{w}_c(0), \boldsymbol{\xi} \rangle| \leq o(\sigma_0 \sigma_\xi).$$

Proof. For any $0 \leq t < T$,

$$\begin{aligned}
& \langle \mathbf{w}_c(t+1), \boldsymbol{\xi} \rangle \\
& = \langle \mathbf{w}_c(t), \boldsymbol{\xi} \rangle + \frac{\eta}{n} \sum_{i=1}^{n} \frac{y^{(i)}}{1 + e^{y^{(i)} F(\mathbf{w}(t), \mathbf{x}^{(i)})}} \psi' \left(\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle \right) \left\langle \boldsymbol{\xi}^{(i)}, \boldsymbol{\xi} \right\rangle \\
& \quad + \frac{\eta}{n} \sum_{i=1}^{n} \left(\frac{1}{1 + e^{y^{(i)} F(\mathbf{w}(t), \mathbf{x}^{(i)})}} \psi' \left(\left\langle \mathbf{w}_c(t), \mathbf{v}_{k_i^*} \right\rangle \right) \left\langle \mathbf{v}_{k_i^*}, \boldsymbol{\xi} \right\rangle \right) \\
& \quad - \frac{\eta}{n} \sum_{i=1}^{n} \left(\frac{1}{1 + e^{y^{(i)} F(\mathbf{w}(t), \mathbf{x}^{(i)})}} \sum_{k \in [K]} \sum_{p \in \mathcal{P}_{bp,k}^{(i)}} \psi' \left(\langle \mathbf{w}_c(t), \alpha_{p,i} \mathbf{v}_k \rangle \right) \langle \alpha_{p,i} \mathbf{v}_k, \boldsymbol{\xi} \rangle \right).
\end{aligned}$$

By Lemma 4, with probability at least $1 - \frac{nK}{\text{poly}d}$, for all $i \in [n]$, $\left\langle \boldsymbol{\xi}^{(i)}, \boldsymbol{\xi} \right\rangle \leq \widetilde{O}(\sigma_\xi^2 d^{-1/2})$ and for all $k \in [K]$, $\langle \mathbf{v}_k, \boldsymbol{\xi} \rangle \leq \widetilde{O}(\sigma_\xi d^{-1/2})$. Then, by $\frac{1}{1+e^{y^{(i)} F(\mathbf{w}(t), \mathbf{x}^{(i)})}} \leq 1$, $\psi'\left(\left\langle \mathbf{w}_c(t), \boldsymbol{\xi}^{(i)} \right\rangle \right) \leq 1$, $\psi'\left(\left\langle \mathbf{w}_c(t), \mathbf{v}_{k_i^*} \right\rangle \right) \leq 1$ and

$\psi'\left(\langle \mathbf{w}_c(t), \alpha_{p,i}\mathbf{v}_k\rangle\right) \le 1$,

$$\begin{aligned}
&|\langle \mathbf{w}_c(t+1), \boldsymbol{\xi}\rangle - \langle \mathbf{w}_c(t), \boldsymbol{\xi}\rangle| \\
&\le \widetilde{O}(\eta\sigma_\xi^2 d^{-1/2}) + \widetilde{O}(\eta\sigma_\xi d^{-1/2}) + \widetilde{O}\left(\eta\alpha^q P\sigma_\xi d^{-1/2}\left(\sigma_0 + \eta T\left(\max_{k'}\rho_{k'} + \sigma_\xi d^{-1/2}\right)\right)^{q-1}\right) \\
&\le \eta\widetilde{O}(\sigma_\xi^2 d^{-1/2} + \sigma_\xi d^{-1/2}).
\end{aligned}$$

The second step uses Lemma 11. The third step uses the upper bound on α. Summing over $0 \le t' \le t$,

$$|\langle \mathbf{w}_c(t), \boldsymbol{\xi}\rangle - \langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle| \le \eta T\widetilde{O}(\sigma_\xi^2 d^{-1/2} + \sigma_\xi d^{-1/2}).$$

When $n \le o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^q d^{1/2}, \sigma_0^{q-1}\sigma_\xi^{q-1} d^{1/2}\right\}\right)$, $K \le o\left(\min\left\{\sigma_0^{q-1}\sigma_\xi^{-1} d^{1/2}, \sigma_0^{q-1} d^{1/2}\right\}\right)$, and $T \le \widetilde{O}\left(\max\left\{n\eta^{-1}\sigma_\xi^{-q}\sigma_0^{-q+2}, K\eta^{-1}\sigma_0^{-q+2}\right\}\right)$,

$$|\langle \mathbf{w}_c(t), \boldsymbol{\xi}\rangle - \langle \mathbf{w}_c(0), \boldsymbol{\xi}\rangle| \le o(\sigma_0\sigma_\xi).$$

□