# Implicit Bias in State Space Models and Linear Autoregressive Training

**Gal Katzhendler**
Faculty of Computer Science
Hebrew University of Jerusalem, Israel
gal.katzhendler@mail.huji.ac.il

**Gilad Yehudai**
Faculty of Computer Science
Faculty of Computer and Electrical Engineering
Technion, Israel

**Gal Vardi**
Department of Computer Science and Applied Mathematics
Weizmann Institute of Science, Israel

## Abstract

Linear State Space Models (SSMs) and linear autoregressive sequence models reuse the same parameters across a sample-dependent number of steps. As a result, different examples impose constraints through different predictors, even though all predictors share the same weights. We study this phenomenon through a general model of multiple non-homogeneous polynomial predictors with shared parameters, in a classification setting. Under standard interpolation and directional-convergence assumptions, together with a trajectory-dependent effective-degree condition, we show that gradient descent on the exponential loss converges in direction to a KKT point of an *effective-degree max-margin problem*. In this problem, only examples with minimal effective degree impose hard margin constraints; examples with faster-growing margins are asymptotically inactive except for feasibility. Our result provides a first-order characterization of the implicit bias in linear SSMs and linear autoregressive models. It exposes a mechanism of length bias: a longer input sequence or additional autoregressive steps may increase the example's effective degree, and the examples with the smallest effective degree dominate the limiting classifier. Our result highlights how non-homogeneity and parameter sharing alter classical homogeneous implicit-bias results.

## 1 Introduction

Ever since the introduction of Transformers (Vaswani *et al.*, 2017), they have been widely used for autoregressive sequence generation, and have come to dominate many areas of machine learning (Chandra *et al.*, 2023; Erol *et al.*, 2024; Ji *et al.*, 2025). Recently, State Space Models (SSMs) have emerged as a promising alternative. By offering improved computational efficiency while maintaining competitive performance, SSMs are increasingly being adopted across a range of tasks (Gu *et al.*, 2021; Gu & Dao, 2023; Ren *et al.*, 2025; Dao & Gu, 2024).

Autoregressive models reuse the same parameters across repeated computation steps. When examples have different sequence lengths or require different numbers of rollout steps, the final prediction for each example may be a different function of the same shared parameters. Even if the per-step computation is linear, the resulting margin is generally a non-homogeneous polynomial in the parameters, with a degree that depends on the number of autoregressive applications.

This creates a setting that is not captured by the existing max-margin implicit-bias theory. Classical results show that gradient descent on separable linear models with an exponential-type loss converges

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

in direction to a hard-margin classifier (Soudry *et al.*, 2018), and later work extends this picture to homogeneous and some non-homogeneous parameterizations (Gunasekar *et al.*, 2018b; Lyu & Li, 2019; Ji & Telgarsky, 2020b; Kunin *et al.*, 2022; Cai *et al.*, 2025). In those settings, however, the training examples are typically evaluated by a single shared predictor, or by predictors with a common homogeneity structure. In contrast, variable-step autoregressive models lead to multiple polynomial predictors, one per example or per length class, all sharing parameters but potentially growing at different asymptotic rates during training.

We study this phenomenon through the lens of shared-parameter, linear-in-the-input polynomial predictors, in a classification setting. Our main observation is that the relevant degree of a predictor is not necessarily its algebraic degree. Along a gradient-descent trajectory, leading homogeneous components may vanish in the limiting direction, causing a high-degree margin polynomial to grow like a lower-degree polynomial. We call this trajectory-dependent quantity the effective degree. Since exponential-loss gradients weight examples by their current margins, examples with larger effective degree can become asymptotically suppressed, while examples with smaller effective degree can determine the limiting classifier.

Our main result gives a conditional max-margin characterization for this setting. Assuming that the loss converges to zero, the parameter direction converges, and each margin has a well-defined integer effective degree along the trajectory, we prove that the limiting direction of gradient descent satisfies the KKT conditions of an effective-degree max-margin problem. The unit-margin constraints are imposed only by examples with minimum effective degree, and higher-effective-degree examples enter only through feasibility constraints. Thus the limiting classifier can be governed by the slowest-growing margins rather than by all examples equally.

We instantiate the theorem in scalar linear autoregressive models, linear state-space models (SSMs), and linear autoregressive SSMs. These models are not intended to capture all features of modern sequence architectures; rather, they isolate the effect of parameter reuse across variable numbers of steps. Indeed, the implicit bias of linear models and linear neural networks has been extensively studied as a testbed for understanding the implicit bias in more complex architectures (Vardi, 2023), and we adopt this approach here by considering linear autoregressive models and linear SSMs. In our simplified settings, when additional recurrent or autoregressive steps increase effective degree, our theorem predicts a form of length-sensitive implicit bias: shorter or lower-effective-degree examples can dominate the asymptotic margin solution.

**Our contribution:**

- We introduce a trajectory-dependent effective degree for non-homogeneous polynomial predictors and show that it can differ sharply from the algebraic degree.
- Under standard asymptotic convergence assumptions and an effective-degree regularity assumption, we prove a max-margin/KKT characterization of the gradient-descent limit for exponential loss. This characterization is significantly different from the standard max-margin characterization of homogeneous models.
- We show that, in the linear autoregressive and SSM instantiations, this characterization yields a length-sensitive bias whenever additional computation steps increase effective degree.

## 2 Related Work

**Implicit bias in homogeneous and near-homogeneous models.** There is a large body of work studying the implicit bias of gradient-based optimization. Soudry *et al.* (2018) initiated this line of work by showing that gradient descent on separable data with the logistic loss converges in direction to the max-margin classifier. Subsequent works extended this perspective to broader classes of models, including linear convolutional networks and homogeneous polynomials (Gunasekar *et al.*, 2018b,a). For homogeneous models, Lyu & Li (2019) showed that any limit point of the optimization path is in the direction of a KKT point of the corresponding margin-constrained Euclidean-norm minimization problem, and Ji & Telgarsky (2020b) established directional convergence of parameters and gradients under additional structural assumptions. Ji & Telgarsky (2020a) used the induced dynamics on dual variables to characterize convergence rates to the implicit bias in linear models; we use a related observation to prove directional convergence of the gradient along a subsequence.

Several works go beyond standard homogeneity. Nacson *et al.* (2019) studied regularization paths for sums of homogeneous models with different degrees, but their setting uses separate parameters for each homogeneous component and does not analyze gradient descent dynamics. Kunin *et al.* (2022) generalized this framework to *quasi-homogeneous* models, where different parameters have different homogeneity degrees, and showed that the resulting implicit bias can be asymmetric. Our polynomial models, as well as their multidimensional extensions, do not fall under this quasi-homogeneous framework; see Appendix A.2. Lippl *et al.* (2022) studied gradient descent on parameter-sharing homogeneous polynomials, allowing the polynomial to vary across examples.

Most closely related to our treatment of non-homogeneous structure is Cai *et al.* (2025), who established implicit bias results for *nearly-homogeneous* models. Although polynomials satisfy their definition of near-homogeneity, their analysis considers a single predictor shared across all data points, whereas we study a more general setting in which each data point may induce a different predictor. Moreover, their result relies on a strong assumption on the optimization dynamics, which we relax for polynomials; see Section 4.2 for a detailed comparison.

**Implicit bias in attention and sequence models.** Recent works have begun to study implicit bias in sequence models. For self-attention, Tarzanagh *et al.* (2023); Ataee Tarzanagh *et al.* (2023); Vasudeva *et al.* (2025) analyze the implicit bias and convergence behavior of attention-based architectures. These works are complementary to ours: they focus on attention mechanisms, whereas our analysis concerns polynomial and state-space/autoregressive structures.

**State-space models and linear autoregressive models.** Several recent theoretical works study state-space models and related linear autoregressive models in settings close to ours. Merrill *et al.* (2025) analyze the expressivity of state-space models. Slutzky *et al.* (2025) study the effect of data poisoning on the implicit bias of diagonal SSMs with the square loss for sequences of fixed length. Other works analyze SSMs with structured transition matrices or related architectures (Gu *et al.*, 2021; Dao & Gu, 2024), while Smékal *et al.* (2024) study the learning dynamics of SSMs under square loss. For autoregressive models, Joshi *et al.* (2025, 2026) establish computational learning results for autoregressive linear threshold models in a setting similar to ours, and Thrampoulidis (2024) study implicit optimization bias in linear autoregressive next-token prediction with fixed-length sequences and cross-entropy loss. Finally, Emami *et al.* (2021) analyze linear RNNs at the NTK limit, showing that the effect of input tokens decays geometrically; the model they consider is reminiscent of the SSMs studied in Section 3.3.

## 3 Preliminaries and Problem Definition

**Notations.** We denote vectors $\mathbf{v} \in \mathbb{R}^d$ by boldface letters i.e. $\mathbf{v} = (v_1, \ldots, v_d)$, and $\bar{\mathbf{v}} := \mathbf{v}/\|\mathbf{v}\|$. We denote the sign function $\operatorname{sgn}(x) := \begin{cases} 1 & x \geq 0 \\ -1 & \text{else} \end{cases}$. We denote by $\|\cdot\|$ the Euclidean norm, and $[n] := \{1, \ldots, n\}$. For $\nu \in \mathbb{N}$, we denote by $\mathcal{P}^\nu$ the sum of (multivariate, vector valued) monomials in $\mathcal{P}$ of degree $\nu$, this is a $\nu$-homogeneous polynomial, which we call the $\nu$-homogenization of $\mathcal{P}$.

### 3.1 Problem Definition

**Model and Data.** We consider a dataset $\{X_i, y_i\}_{i=1}^N$ where $X_i \in \mathbb{R}^{d \times l_i}$ and $y_i \in \{\pm 1\}$. We view each input sample $X_i$ as $l_i$ tokens, each of dimension $d$. Note that we allow for each sample to have a different number of tokens. The difficulty in determining the implicit bias of our models already appears in the case of $d = 1$. Hence, for brevity, in the main part of the paper, we consider the case of $d = 1$, namely, each token is a scalar in $\mathbb{R}$, and we denote the input sequence as $\mathbf{x}_i \in \mathbb{R}^{l_i}$. All our results are readily extended to the case where $d$ is some constant larger than 1, and we prove these extensions formally in Appendix B.

Let $\{\mathbf{x}_i, y_i\}_{i=1}^N$ be a dataset where $\mathbf{x}_i \in \mathbb{R}^{l_i}$, $y_i \in \{\pm 1\}$. We define a set of models $\{\mathcal{M}_i : \mathbb{R}^D \to \mathbb{R}^{l_i}\}_{i=1}^N$, that share the same parameters $\mathbf{w} \in \mathbb{R}^D$, but might have different output dimensions. These models act linearly on the data $\mathbf{x}_i$, namely their output is $\hat{y}_i := \langle \mathcal{M}_i(\mathbf{w}), \mathbf{x}_i \rangle$ where $l_i$ is the dimension of the $i$-th sample, i.e. $\mathbf{x}_i \in \mathbb{R}^{l_i}$, or in other words, the number of one dimensional tokens $\mathbf{x}_i$ has. Note that although $\{\mathcal{M}_i : \mathbb{R}^D \to \mathbb{R}^{l_i}\}_{i=1}^N$ indexes $N$ models, they may not all be distinct, i.e., we allow $\mathcal{M}_i = \mathcal{M}_j$ for some $i \neq j$. This multi-model setting is a generalization of the standard implicit bias

setting where all the samples share the same predictor $\mathcal{M}$. This framing aligns with the setting of SSMs and autoregressive models, where each sample may consist of a different number of input tokens or autoregressive steps. In this paper, we focus on two such models: (1) a linear autoregressive model and (2) a linear state-space model (SSM), formally defined in Sections 3.2 and 3.3. Although the models acts linearly on the input data, their parameterizations can be highly non-linear, leading to implicit biases that differ substantially from those of standard linear predictors.

**Training.** In this paper we will consider the exponential loss and denote $\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} \exp(-y_i \mathcal{M}_i(\mathbf{w})^\top \mathbf{x}_i)$. Our results can be readily extended to the logistic loss as well. Throughout the paper, we consider gradient descent w.r.t. vectors or matrices of the parameters. We denote by $\mathbf{w}^{(0)}$ the parameters at initialization, and

$$\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)} - \eta_t \nabla \mathcal{L}(\mathbf{w}^{(t-1)}) \,, \tag{1}$$

as the parameter vector after the $t^{\text{th}}$ step of gradient descent, where $\eta_t$ denotes the learning rate at time $t$. Our results can be applied to general learning rate schedulers, and we only assume the learning rate sequence $\{\eta_t\}_{t=1}^{\infty}$ does not increase. We denote by $\bar{\mathbf{w}}^{\infty} = \lim_{t\to\infty} \frac{\mathbf{w}^{(t)}}{\|\mathbf{w}^{(t)}\|}$ the directional limit of $\mathbf{w}$ when it exists.

### 3.2 Linear Autoregressive Model

The first model we consider is a linear autoregressive model. This type of models is common in many academic fields for the prediction of time series Gujarati & Porter (2009); Wei (2013); Kirchgässner *et al.* (2012); Dama & Sinoquet (2021); Chen *et al.* (2025). Moreover, it is a basic simplification of large language models, where we isolate the autoregressive part of transformers from the non-linear parts which are the self-attention and MLP. The core mechanism is that the model generates new tokens, and then apply itself autoregressively on its own output.

We consider an example $\mathbf{x} \in \mathbb{R}^l$ which we interpret as an input prompt, i.e., the initial part of some sequence. Here we will consider parameters $\mathbf{w} \in \mathbb{R}^D$ and assume for simplicity $D = l$, thus $\mathbf{x}^\top = \mathbf{x}_{1:D}^\top = (x_1, \ldots, x_D)$. We compute a sequence of $T$ autoregressive steps, where the final token's sign is the label. Formally, the model computes:

$$x_{D+1} = \langle \mathbf{w}, \mathbf{x} \rangle \,, \quad x_{D+2} = \langle \mathbf{w}, \mathbf{x}_{2:D+1} \rangle \,, \quad \ldots \quad , x_{D+T} = \langle \mathbf{w}, \mathbf{x}_{T:D+T-1} \rangle \,. \tag{2}$$

The label of the sequence is defined as $\operatorname{sgn}(x_{D+T})$. More generally, we may allow $l \neq D$: in the case $l < D$, we pad $\mathbf{x} = (0, \ldots, x_1, \ldots, x_l)$, and in the case $l > D$, apply $\mathbf{w}$ only to the last $D$ elements of $\mathbf{x}$. This can be thought of as the input prompt being longer than the model's context length.

We denote the vector of coefficients of $x_1, \ldots, x_D$ in $x_{D+T}$ as $\mathcal{P}$, a polynomial in $\mathbf{w}$. Using this notation, we have that $\langle \mathcal{P}(\mathbf{w}), \mathbf{x} \rangle = x_{D+T}$. See Figure 1 for a demonstration.

We can interpret the setup in the context of language models as having $l$ initial one-dimensional tokens, $x_1, \ldots, x_l$, which serve as the starting sequence. In standard transformers, and autoregressive language models in general, text is generated token by token: the model predicts the next token based on the current sequence, appends this token, and then feeds the entire updated sequence back into the model to produce the next one, $T$ times. Our framework mirrors this process but uses a linear autoregressive model, which is more tractable to analyze. We note that in our work the polynomials are usually associated with an indexed example $\mathbf{x}_i \in \mathbb{R}^{l_i}$, hence we denote them $\mathcal{P}_i$. Note that crucially for a dataset $\{\mathbf{x}_i, y_i\}_{n=1}^{N}$, we have a number of autoregressive steps $T_i$ for each example, which together with $l_i$ changes $\mathcal{P}_i$. That is, both elements embodied in the polynomial predictor. This also means that we may have $\mathcal{P}_i = \mathcal{P}_j$ for $i \neq j$, if both examples $i, j$ are labeled using the same number $T_i$ of autoregressive steps and share the same input length $l_i$. We note that in this model our training setup corresponds to training directly on the output after $T_i$ steps, rather than training on the next-token-prediction.

### 3.3 Linear State Space Models

The second model we consider, is a Linear State Space Model (SSM). SSMs are used to model natural phenomena and have been lately adopted by the Machine Learning community as a promising candidate for Large Language Models to replace and complement Transformers (Gu *et al.*, 2021; Gu & Dao, 2023; Dao & Gu, 2024).

We consider a sequence of tokens of variable lengths $X \in \mathbb{R}^{d \times l}$ where $l$ is the length of the sequence. SSMs consist of parameter matrices $A \in \mathbb{R}^{d_A \times d_A}$, $B \in \mathbb{R}^{d_A \times d}$, $C \in \mathbb{R}^{d_A \times d_C}$ where the model is given recursively by a sequence of "hidden" states $\mathbf{h}$, i.e., for $1 \le i \le l$, $\mathbf{h}_i = A \cdot \mathbf{h}_{i-1} + B(X)_i$, where $(X)_i$ denotes the $i^{\text{th}}$-column of $X$, with $\mathbf{h}_0 = \mathbf{0}$, and with the output being $y = \text{sgn}(C^\top \mathbf{h}_l)$. Note that we consider binary classification and assume $d_C = 1$. For the rest of the main part of the paper, we consider the case $d = 1$, and let $B, C$ be vectors. We provide the generalization for a matrix $C$ and full proof for the more general case of $d \ge 1$ in Appendix B.

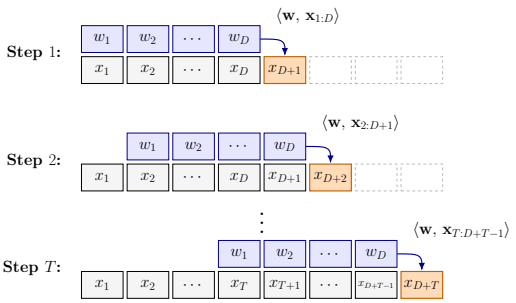


Figure 1: Movie-strip view of the autoregressive construction.

A useful convolutional way to denote it overall is

$$y = \text{sgn}\left( C^\top \sum_{i=1}^{l} A^{l-i} B(X)_i \right) .$$

Note that SSMs in practice usually use nonlinearities, which we omit, and sometimes use special matrix structure for $A$, such as diagonal, or diagonal up to some low rank. We analyze the general case and note that in principle our analysis should apply to those cases as well.

With a dataset $\{\mathbf{x}_i, y_i\}_{n=1}^N$, denote the number of tokens of the $k^{\text{th}}$ sample as $l_k$, and consider the polynomial

$$Q_k(A, B, C) = \begin{pmatrix} C^\top A^{l_k - 1} B \\ \vdots \\ C^\top AB \\ C^\top B \end{pmatrix}$$

i.e. $y_k = \langle Q_k(A, B, C), \mathbf{x}_k \rangle$, note that $Q_k$ depends only on the number of steps taken by the SSM, that is the model works for a variable length input $\mathbf{x}_k \in \mathbb{R}^{l_k}$. Similarly to the linear autoregressive case, although each sample $k$ is associated with a different polynomial $Q_k$, we may have $Q_k = Q_j$ for $k \ne j$ if both examples $k, j$ have the same length, and $l_k$ is embodied in $Q_k$.

### 3.4 Linear Autoregressive SSM

Note that the SSM from the previous section does not use its own outputs as inputs for the next steps. That is, the SSM simply runs over the input sequence but does not extend it. Hence it is not autoregressive in this sense. We now introduce the *Autoregressive SSM*. Consider an example $\mathbf{x} \in \mathbb{R}^l$, that we can think of as an input prompt, or as a beginning of some sequence $\mathbf{x}^\top = \mathbf{x}_{1:l}^\top = (x_1, \ldots, x_l)$, we compute a sequence of $T$ autoregressive steps:

$$x_{l+1} = C^\top \sum_{i=1}^{l} A^{l-i} B x_i, \quad x_{l+2} = C^\top \sum_{i=1}^{l+1} A^{l+1-i} B x_i, \quad \ldots \quad , x_{l+T} = C^\top \sum_{i=1}^{l+T-1} A^{l+T-1-i} B x_i \, . \tag{3}$$

The label of the sequence $\mathbf{x}$ is defined as $\text{sgn}(x_{l+T})$, and we write the polynomial coefficients as a vector $Q(A, B, C)$, i.e. $y = \text{sgn}\left(\langle Q(A, B, C), \mathbf{x} \rangle\right)$. Note that $Q$ depends both on the number of autoregressive steps $T$ *and* the number of input tokens $l$. As in previous sections, $Q_i$ implicitly embeds $l_i$ and $T_i$, and we allow $Q_i = Q_j$ for $i \ne j$.

## 4 Assumptions on the Dynamics

In this section, we present the assumptions about the optimization dynamics that we will use to prove our main results. Before introducing our assumptions, we define the "effective degree" of a polynomial over an optimization trajectory:

**Definition 4.1** (Effective Degree)**.** *Let $p : \mathbb{R}^D \to \mathbb{R}$ be a polynomial with $\deg(p) = k$. Let $\mathbf{w}^{(t)} \in \mathbb{R}^D$ be some trajectory of the polynomial's inputs, such that $\left\|\mathbf{w}^{(t)}\right\| \to \infty$ and $\lim_t \frac{\mathbf{w}^{(t)}}{\left\|\mathbf{w}^{(t)}\right\|} = \bar{\mathbf{w}}^\infty$. We say*

*$p$ has an* effective degree *of $m$ over the trajectory $\mathbf{w}^{(t)}$ (denoted* $\text{effdeg}_{\mathbf{w}^{(t)}}(p)$*, or simply* $\text{effdeg}(p)$ *when $\mathbf{w}^{(t)}$ is implicit) if*

$$\lim_{t\to\infty} \frac{p(\mathbf{w}^{(t)})}{\left\|\mathbf{w}^{(t)}\right\|^j} = \begin{cases} Cp^j(\bar{\mathbf{w}}^\infty) \neq 0 & j = m \\ 0 & j > m \end{cases},$$

*and similarly*

$$\lim_{t\to\infty} \frac{\nabla p(\mathbf{w}^{(t)})}{\left\|\mathbf{w}^{(t)}\right\|^{j-1}} = \begin{cases} C\nabla p^j(\bar{\mathbf{w}}^\infty) \neq \mathbf{0} & j = m \\ \mathbf{0} & j > m \end{cases}.$$

*For some constant $C > 0$.*

Intuitively, the definition states that along the trajectory, $\mathbf{w}^{(t)}$ moves in a direction that cancels out higher degree monomials (if such monomials exist). A simple example would be the trajectory $\mathbf{w}^{(t)} = (t, t, t)$ with the polynomial $p(\mathbf{w}) = w_1 + (w_2 - w_3)^2$, then $p(\mathbf{w}^{(t)}) = t = w_i$ for all $i$. From now on, for a polynomial $p$, we refer to $\deg(p)$ as the algebraic degree of $p$, as to not confuse it with its effective degree.

We use the following assumptions throughout the paper.

**Assumption 1** (Loss minimization)**.** *The iterates $\mathbf{w}^{(t)}$ minimize the objective, i.e. $\mathcal{L}(\mathbf{w}^{(t)}) \underset{t\to\infty}{\to} 0$.*

**Assumption 2** (Convergence in direction)**.** *The model's parameters converge in direction, i.e.* $\lim_{t\to\infty} \frac{\mathbf{w}^{(t)}}{\left\|\mathbf{w}^{(t)}\right\|} = \bar{\mathbf{w}}^\infty$.

**Assumption 3.** *$\forall i$,* $\text{effdeg}_{\mathbf{w}^{(t)}}(\mathcal{P}_i^\top \mathbf{x}_i y_i) = \nu_i$ *for $\nu_i \in \mathbb{N}$*[1]*, with some $C_i > 0$.*

In our first assumption, we assume the iterates of the parameters $\mathbf{w}^{(t)}$ minimize the objective, i.e. we assume that the loss converges to zero. This is a practical assumption as in many cases deep learning exhibits complete interpolation i.e. zero or close to zero error on the data. It also allows us to separate the question of whether gradient descent interpolates from the question of where gradient descent converges when it interpolates. As such, this assumption is also common in literature (e.g., Gunasekar *et al.* (2018b,a); Vardi & Shamir (2021); Timor *et al.* (2022)).

Our second assumption states that the parameter vector $\mathbf{w}^{(t)}$ converges in direction. This assumption means that the trajectory of $\mathbf{w}^{(t)}$ is "well behaved", and does not oscillate indefinitely. This is a standard assumption in the literature, and although some works, such as Ji & Telgarsky (2020b); Cai *et al.* (2025) have managed to prove it under some assumptions, many works use it implicitly or explicitly as an assumption (e.g., Kunin *et al.* (2022); Gunasekar *et al.* (2018b); Lyu & Li (2019)). Studying when directional convergence can be proved in our setting is a natural and important open question; we leave it to future work and focus here on characterizing the limit.

Our third assumption is more elaborate, but it is natural when we think of polynomials. If we fix some unit vector $\bar{\mathbf{w}}^\infty$, then for any polynomial $p(\mathbf{w})$, the univariate polynomial $q\left(\left\|\mathbf{w}^{(t)}\right\|\right) := p\left(\left\|\mathbf{w}^{(t)}\right\| \bar{\mathbf{w}}^\infty\right) = \Theta\left(p^k\left(\left\|\mathbf{w}^{(t)}\right\| \bar{\mathbf{w}}^\infty\right)\right)$ when $\left\|\mathbf{w}^{(t)}\right\| \to \infty$, for the maximal $k$ for which $p^k(\bar{\mathbf{w}}^\infty) \neq 0$. In words, when we plug in a fixed direction to a multivariate polynomial, it starts behaving like a univariate polynomial, hence grows with its degree. In this case, its degree would be the highest degree that does not cancel when plugging the direction in. The gradient part of the definition ensures that the optimization does not go too far off this rather intuitive behavior.

### 4.1 Initialization-Dependent Effective Degree

Having the effective degree dependent on the gradient descent trajectory may seem like a burdensome complication, which makes the calculation of the implicit bias highly contextual. We show that under the setting of our non-homogeneous polynomials, the effective degree may be different for initializations in two different open sets. Hence, instead of only depending on the model and training examples, it depends on the initialization as well.

[1] We note that a polynomial with $\text{effdeg}_{\mathbf{w}^{(t)}}(\mathcal{P}^\top \mathbf{x} y) = 0$ contradicts Assumption 1, hence for the polynomials we will consider, effdeg $\geq 1$.

**Lemma 4.2.** *There exist a polynomial predictor $\mathcal{P} : \mathbb{R}^D \to \mathbb{R}^n$, a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^n$ and two disjoint open sets $U_2, U_4 \subset \mathbb{R}^D$, such that when trained with gradient descent (Equation (1)) with a constant step size $\eta > 0$, the margins $m_i = y_i \langle \mathcal{P}, \mathbf{x}_i \rangle$ have:*

$$\operatorname{effdeg}(m_i) = \begin{cases} 2i & \mathbf{w}^{(0)} \in U_2 \\ 4i & \mathbf{w}^{(0)} \in U_4 \end{cases}$$

The full proof is in Appendix D.1. The proof holds for every $n$, and for any subset of the above dataset, and it can be extended to a larger family of predictors beyond that in the lemma. The example above requires careful analysis of the optimization trajectory, but this is not always the case. We provide a concrete, nontrivial family of predictors and datasets, for which the effective degree follows from a short calculation, in Appendix D.2.

### 4.2 Comparison with Cai *et al.* (2025)

At this point, one might think that the implicit bias in our setting can be analyzed using the results from Cai *et al.* (2025), at least for the case where all examples share the same predictor. Cai *et al.* (2025) established a result on the implicit bias for the class of *nearly-homogeneous models*, which include our polynomial setting. Perhaps surprisingly, their result is restrictive, as the assumptions of Cai *et al.* (2025) implicitly imply that the effective degree of the model equals its algebraic degree. Thus, we can think of Assumption 3 as a relaxation which captures a broader variety of dynamical behaviors.

**Assumption 4** (Cai *et al.* (2025), simplified for our setting)**.** *At some point $t_0$ during optimization, $\mathcal{L}(\mathbf{w}^{(t_0)}) \leq \exp\left(-p\left(\left\|\mathbf{w}^{(t_0)}\right\|\right)\right)$ for some $p(x)$. If the predictor is a polynomial of the form $f(\mathbf{w}; \mathbf{x}) = \langle q(\mathbf{w}), \mathbf{x} \rangle$, then $0 \leq \deg(p) \leq \deg(q) - 1$, where for the special case of an homogeneous $q$, $p \equiv 0$*

Assumption 4 is presented in Cai *et al.* (2025) as a strong separation assumption on the data. However, it is essentially an assumption on the optimization rate. Since it includes $t_0$ on both sides of the inequality, it requires that at some point, the optimization reaches a certain rate; specifically, it requires that at some point all margins behave as if they were some polynomial w.r.t. the model's parameters. The problem is that the condition implies that all margins progress at some maximal rate. Is it possible that this maximal rate is never reached, and hence Assumption 4 does not hold? As we shall see next, the answer is positive even for a very simple setting with our linear autoregressive model.

**Lemma 4.3.** *Consider the linear autoregressive setting from Section 3.2, trained using gradient descent (Equation (1)) with a constant step size $\eta > 0$ and initialization at $\mathbf{0}$. Then, there exists a size-1 dataset $\{(\mathbf{x}, y)\}$, such that Assumption 4 from Cai* et al. *(2025) does not hold, while our Assumptions 1 to 3 hold.*

The example we provide has a margin (polynomial w.r.t. $\mathbf{w}$), with algebraic degree 4, but an effective degree of 1. This is in contradiction to Assumption 4 that implies the effective degree is higher. Moreover, if we do not insist on the setting from Section 3.2, we can find such examples for every even algebraic degree $d$, and $k < d$ effective degree (see Appendix C.3), which in the case of even $k$ work for almost all initializations. Therefore, a more nuanced approach is necessary to capture the range of possible dynamics of optimization, Assumption 3 is one such approach. In fact, in the following lemma we prove that in the case of polynomials, Assumption 4 implies Assumptions 1 to 3 (see Appendix A.4 for the proof):

**Lemma 4.4.** *In the setting where the predictor is polynomial, Assumption 4 implies Assumptions 1 to 3.*

Assumption 4 essentially requires that the effective degree for every example is equal to the algebraic degree, while Assumption 3 allows *any* effective degree. As we will see in Section 5, the implicit bias of gradient descent is different than the one presented in Cai *et al.* (2025) for the cases where Assumption 4 does not hold. In Section 6, we empirically verify that our assumptions (including Assumption 3) indeed hold, at least in some cases.

We further emphasize that while the above discussion shows that for the case where all examples share the same predictor $\mathcal{P}_i$ (i.e., the same $l_i$ and $T_i$), our Assumption 3 is weaker than Assumption 4 of Cai *et al.* (2025), cases where $l_i$ or $T_i$ (and hence $\mathcal{P}_i$) are not uniform for all examples are not covered by the setting in Cai *et al.* (2025) at all. On the other hand, Cai *et al.* (2025) cover non-polynomial model, which we do not consider.

Finally, we note that it is theoretically possible that the dynamics do not satisfy our assumptions. This can happen when the effective degree of the polynomial along the trajectory behaves as a "fractional-degree". We elaborate on these cases in Appendix A.3. Studying such dynamics introduces a challenging analysis since the margin in these cases, as well as the KKT conditions, is not well-defined. We leave investigating those cases for future work.

## 5 Main Results

We will now present our main result, and show how it applies to the various models we consider.

**Theorem 5.1.** *Let $\{\mathcal{P}_i\}_{i=1}^N$ a collection of polynomials with parameters $\mathbf{w} \in \mathbb{R}^D$. For a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^N$ where $y_i = \operatorname{sgn}(\langle \mathcal{P}_i(\mathbf{w}), \mathbf{x}_i \rangle)$, consider the sequence of gradient updates (Equation (1)) with a non-increasing sequence of step sizes $\{\eta_t\}_t$, minimizing the empirical risk. If Assumptions 1 to 3 hold, we have that $\bar{\mathbf{w}}^\infty := \lim_{t\to\infty} \frac{\mathbf{w}^{(t)}}{\|\mathbf{w}^{(t)}\|}$ is a positive scaling of a first order stationary point (KKT point) of the following optimization problem:*

$$\min_{\mathbf{w}\in\mathbb{R}^D} \|\mathbf{w}\|_2^2 \ \text{ s.t. } \forall i, y_i \left\langle \mathbf{x}_i, \mathcal{P}_i^{\nu_i}(\mathbf{w}) \right\rangle \geq 0 \qquad \text{(OPT)}$$
$$\forall i \in Z, y_i C_i \left\langle \mathbf{x}_i, \mathcal{P}_i^{\nu_i}(\mathbf{w}) \right\rangle \geq 1 \,,$$

*where $Z = \arg\min_{i\in[n]} \left\{ \operatorname{effdeg}_{\mathbf{w}} y_i \langle \mathcal{P}_i, \mathbf{x}_i \rangle \right\}$.*

The idea behind the proof is to establish the dominance of the lowest effective-degree margins on the gradient, as higher effective degrees decay faster. Then, analyzing the induced gradient-descent updates on the dual weights, i.e., the dual certificate, to show a subsequence of these weights converges. Finally, we use a convexity argument to show the gradient under this subsequence converges in direction to $\bar{\mathbf{w}}^\infty$. An extended proof intuition and the full proof are in Appendix E.

Note that this result is different in numerous ways from the implicit bias of homogeneous models, given by

$$\min_{\mathbf{w}} \|\mathbf{w}\| \ \text{ s.t. } \forall i, \ y_i f(\mathbf{w}) \geq 1 \,.$$

1. The positive-margin constraint is only w.r.t. some of the samples, specifically the ones with lowest effective degree. The constraint on the rest of the samples only roughly requires that they will be classified correctly.
2. Each sample is associated with a different predictor.
3. Importantly, the margins $y_i C_i \left\langle \mathbf{x}_i, \mathcal{P}_i^{\nu_i}(\mathbf{w}) \right\rangle$ use different predictors than those being optimized ($\mathcal{P}_i$ rather than $\mathcal{P}_i^{\nu_i}$), that is, the direction of convergence is towards a KKT point of an optimization problem which may a-priori seem unrelated to the optimized objective. This is similar to Cai *et al.* (2025); however, they use the algebraic-degree homogenization of the optimized predictors instead of the effective-degree homogenization.
4. The implicit bias is different depending on the dynamics of gradient descent and depends on the effective degree, which is optimization dependent. This is similar in spirit to Ataee Tarzanagh *et al.* (2023); Tarzanagh *et al.* (2023), where multiple implicit biases are possible depending on the optimization dynamics.

Note that Theorem 5.1 is phrased in a rather general sense for polynomial predictors, but since we know each of our models is in fact a polynomial, we can see how it applies to them.

**Corollary 5.2.** *Under the settings of Theorem 5.1, $\bar{\mathbf{w}}^\infty$ is a positive scaling of a first order stationary point of Equation (OPT), where:*

- *For linear autoregressive models, the polynomial predictors are defined as in Section 3.2.*
- *For linear SSMs, the polynomial predictors are defined as in Section 3.3.*
- *For linear autoregressive SSMs, the polynomial predictors are defined as in Section 3.4.*

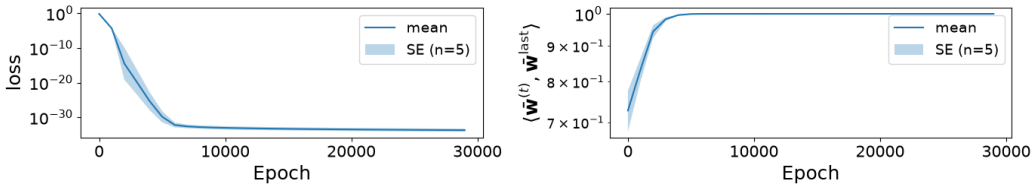


Figure 2: **Left:** The loss. **Right:** Cosine similarity to the final predictor direction. Y-axes are log-scale.

**Interpretation of the max-margin solution.** We split the interpretation into two cases: (1) All polynomials have their maximal possible effective degree, i.e. $\forall i, \deg(\mathcal{P}_i) = \text{effdeg}(\mathcal{P}_i)$. (2) Some polynomials do not achieve their maximal effective degree, i.e. $\exists i, \deg(\mathcal{P}_i) > \text{effdeg}(\mathcal{P}_i)$.

In the first case, our results imply that the implicit bias for a max-margin solution is affected mainly by the shortest sequences seen during training (i.e. polynomials with lowest degree). The other sequences are only required to obtain a correct classification. In the second case, the solution is mainly affected by the examples with the lowest effective degree. This very roughly extends the interpretation of the first case, in the sense that fewer autoregressive steps, or lower-degree polynomials, are most likely to have a lower effective degree. For example, in the case of polynomials corresponding to only one autoregressive step, these polynomials are *guaranteed* to have the minimal effective degree, and hence have greater effect on the max-margin.

## 6 Experiments

The discussion in Section 4.2 alludes that our assumptions may not hold in general, hence it's customary to check them empirically. We performed experiments with autoregressive linear models and SSMs both on synthetic data and on MNIST (LeCun, 1998), and found many cases for which our assumptions hold. In the main text, we present our experiments on autoregressive linear models with a synthetic teacher-student dataset. Here, we generate a random teacher and use it to label Gaussian examples, with different numbers of autoregressive steps $T \in \{1, 3, 5\}$. We refer the reader to Appendix G for experiments with SSMs and linear autoregressive models in other settings, and for more information about the experiments.

### 6.1 Verifying Our Assumptions

In Figures 2 and 3 we verify our assumptions. First, we plot the loss to verify Assumption 1. We see that the loss approaches 0. Next, to verify Assumption 2 we consider the correlation of the parameter direction with the final predictor, i.e. $\left\langle \bar{\mathbf{w}}^{(t)}, \bar{\mathbf{w}}^{\infty} \right\rangle$. Here, since we do not have access to the directional limit $\bar{\mathbf{w}}^{\infty}$, we use the last iterate instead. We see that iterates become close to this direction quite early in training. For Assumption 3, we plot the following ratios for each training example. We plot $\left| \frac{m_i^{\text{effdeg}(m_i)}(\mathbf{w}^{(t)})}{m_i(\mathbf{w}^{(t)})} \right|$, likewise $\frac{\left\| \nabla m_i^{\text{effdeg}(m_i)}(\mathbf{w}^{(t)}) \right\|}{\left\| \nabla m_i(\mathbf{w}^{(t)}) \right\|}$, and $\frac{\left| m_i^{\text{effdeg}(m_i)}(\mathbf{w}^{(t)}) / m_i(\mathbf{w}^{(t)}) \right|}{\left\| \nabla m_i^{\text{effdeg}(m_i)}(\mathbf{w}^{(t)}) \right\| / \left\| \nabla m_i(\mathbf{w}^{(t)}) \right\|}$ where $m_i(\mathbf{w}^{(t)}) := y_i \left\langle \mathcal{P}_i(\mathbf{w}^{(t)}), \mathbf{x}_i \right\rangle$. Assumption 3 requires the first and second ratios to converge to the same constants, with the third ratio converging to 1. For details about finding the effective degree, see Appendix G.

### 6.2 Convergence to Max-Margin

In Figure 4 we plot the minimum margins per effective degree. We see that margins of examples with the minimal effective degree indeed converge to higher values, in accordance to the constraints of Equation (OPT) and the results of Theorem 5.1. We also plot the outputs per effective degree after training. This verifies that examples with lower effective degree have margins bounded away from zero, and shows the concentration of margins in higher-effective degrees, which matches Equation (OPT).

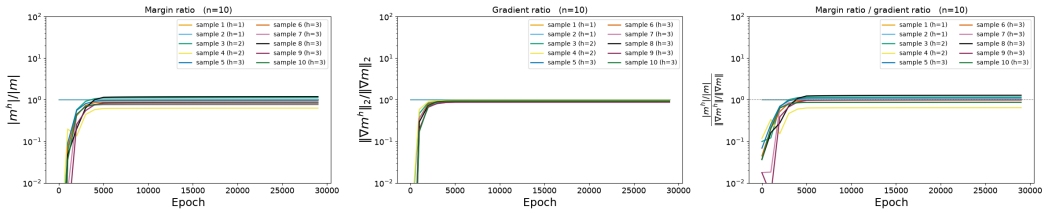


Figure 3: **Left:** Ratio between the effective degree homogenization of the margin and the margin, per example. **Middle:** The corresponding ratio for the norms of the gradients of the homogenization of the margin, and the margin. **Right:** The ratio of the two former ratios. Y-axes are log-scale.

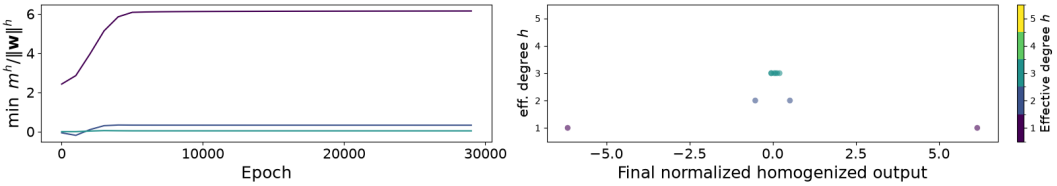


Figure 4: **Left:** Minimal normalized, homogenized margin per effective degree. **Right:** All points with their effective degrees and final normalized homogenized margins.

## 7 Discussion and future work

In this work, we characterized the implicit bias of gradient descent for polynomial predictors where each sample is labeled by a different predictor. It implies implicit bias results for Linear Autoregressive Models and linear SSMs, including linear Autoregressive SSMs. Finally, we shed light on the implicit bias of non-homogeneous models through the lens of polynomials, demonstrating that previous assumptions about their dynamics (Cai *et al.* (2025)) may be overly strong in certain settings.

Some interesting questions remain for future work: First, can we completely characterize the implicit bias for the multi-polynomial case? We found that sometimes, the effective degree may be fractional (Appendix A.3). What is the implicit bias in those cases? Second, how can we generalize these polynomial results to general non-homogeneous models? We consider only polynomials in this paper, but similar results may apply to other classes of predictors and be of much interest.

### Acknowledgements

GK is funded by the European Research Council (ERC) under the European Union's Horizon 2022 research and innovation program (grant agreement No. 101041711), and the Simons Foundation (as part of the Collaboration on the Mathematical and Scientific Foundations of Deep Learning). GV is supported by the Israel Science Foundation (grant No. 2574/25), by a research grant from Mortimer Zuckerman (the Zuckerman STEM Leadership Program), and by research grants from the Center for New Scientists at the Weizmann Institute of Science, and the Shimon and Golde Picker – Weizmann Annual Grant.

## References


Ataee Tarzanagh, Davoud, Li, Yingcong, Zhang, Xuechen, & Oymak, Samet. 2023. Max-margin token selection in attention mechanism. *Advances in neural information processing systems*, **36**, 48314–48362.

Cai, Yuhang, Zhou, Kangjie, Wu, Jingfeng, Mei, Song, Lindsey, Michael, & Bartlett, Peter. 2025. Implicit Bias of Gradient Descent for Non-Homogeneous Deep Networks. *Pages 6426–6504 of:* Singh, Aarti, Fazel, Maryam, Hsu, Daniel, Lacoste-Julien, Simon, Berkenkamp, Felix, Maharaj, Tegan, Wagstaff, Kiri, & Zhu, Jerry (eds), *Proceedings of the 42nd International Conference on Machine Learning*. Proceedings of Machine Learning Research, vol. 267. PMLR.

Chandra, Abel, Tünnermann, Laura, Löfstedt, Tommy, & Gratz, Regina. 2023. Transformer-based deep learning for predicting protein properties in the life sciences. *eLife*, **12**(jan), e82819.

Chen, Xinyu, Digalakis, V., Ding, L., Zhuang, Dingyi, & Zhao, Jinhua. 2025. Interpretable Time Series Autoregression for Periodicity Quantification. *ArXiv*, **abs/2506.22895**.

Dama, F., & Sinoquet, Christine. 2021. Partially Hidden Markov Chain Linear Autoregressive model: inference and forecasting. *ArXiv*, **abs/2102.12584**.

Dao, Tri, & Gu, Albert. 2024. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*.

Emami, Melikasadat, Sahraee-Ardakan, Mojtaba, Pandit, Parthe, Rangan, Sundeep, & Fletcher, Alyson K. 2021. *Implicit Bias of Linear RNNs*.

Erol, Mehmet Hamza, Senocak, Arda, Feng, Jiu, & Chung, Joon Son. 2024. Audio Mamba: Bidirectional State Space Model for Audio Representation Learning. *IEEE Signal Processing Letters*, **31**, 2975–2979.

Gu, Albert, & Dao, Tri. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, Albert, Goel, Karan, & Ré, Christopher. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Gujarati, Damodar N, & Porter, Dawn C. 2009. *Basic econometrics*. McGraw-hill.

Gunasekar, Suriya, Lee, Jason, Soudry, Daniel, & Srebro, Nathan. 2018a. Characterizing implicit bias in terms of optimization geometry. *Pages 1832–1841 of: International Conference on Machine Learning*. PMLR.

Gunasekar, Suriya, Lee, Jason D, Soudry, Daniel, & Srebro, Nati. 2018b. Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, **31**.

Ji, Ke, Jiahao Xu, Liang, Tian, Qiuzhi Liu, Zhiwei He, Xingyu Chen, Xiaoyuan Liu, Zhijie Wang, Junying Chen, Benyou Wang, Zhaopeng Tu, Haitao Mi, & Yu, Dong. 2025. The First Few Tokens Are All You Need: An Efficient and Effective Unsupervised Prefix Fine-Tuning Method for Reasoning Models.

Ji, Ziwei, & Telgarsky, Matus. 2020a. *Characterizing the implicit bias via a primal-dual analysis*.

Ji, Ziwei, & Telgarsky, Matus. 2020b. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, **33**, 17176–17186.

Joshi, Nirmit, Vardi, Gal, Block, Adam, Goel, Surbhi, Li, Zhiyuan, Misiakiewicz, Theodor, & Srebro, Nathan. 2025. *A Theory of Learning with Autoregressive Chain of Thought*.

Joshi, Nirmit, Magen, Roey, Srebro, Nathan, Tsilivis, Nikolaos, & Vardi, Gal. 2026. *Learning to Think from Multiple Thinkers*.

Kirchgässner, Gebhard, Wolters, Jürgen, & Hassler, Uwe. 2012. *Introduction to modern time series analysis*. Springer Science & Business Media.

Kunin, Daniel, Yamamura, Atsushi, Ma, Chao, & Ganguli, Surya. 2022. The asymmetric maximum margin bias of quasi-homogeneous neural networks. *arXiv preprint arXiv:2210.03820*.

LeCun, Yann. 1998. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Lippl, Samuel, Abbott, L. F., & Chung, SueYeon. 2022. *The Implicit Bias of Gradient Descent on Generalized Gated Linear Networks*.

Lyu, Kaifeng, & Li, Jian. 2019. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*.

Merrill, William, Petty, Jackson, & Sabharwal, Ashish. 2025. *The Illusion of State in State-Space Models*.

Nacson, Mor Shpigel, Gunasekar, Suriya, Lee, Jason D., Srebro, Nathan, & Soudry, Daniel. 2019. *Lexicographic and Depth-Sensitive Margins in Homogeneous and Non-Homogeneous Deep Models*.

Paszke, Adam, Gross, Sam, Massa, Francisco, Lerer, Adam, Bradbury, James, Chanan, Gregory, Killeen, Trevor, Lin, Zeming, Gimelshein, Natalia, Antiga, Luca, *et al.* 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, **32**.

Ren, Liliang, Liu, Yang, Lu, Yadong, Shen, Yelong, Liang, Chen, & Chen, Weizhu. 2025. *Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling*.

Slutzky, Yonatan, Alexander, Yotam, Razin, Noam, & Cohen, Nadav. 2025. *The Implicit Bias of Structured State Space Models Can Be Poisoned With Clean Labels*.

Smékal, Jakub, Smith, Jimmy T. H., Kleinman, Michael, Biderman, Dan, & Linderman, Scott W. 2024. *Towards a theory of learning dynamics in deep state space models*.

Soudry, Daniel, Hoffer, Elad, Nacson, Mor Shpigel, Gunasekar, Suriya, & Srebro, Nathan. 2018. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, **19**(70), 1–57.

Tarzanagh, Davoud Ataee, Li, Yingcong, Thrampoulidis, Christos, & Oymak, Samet. 2023. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*.

Thrampoulidis, Christos. 2024. *Implicit Optimization Bias of Next-Token Prediction in Linear Models*.

Timor, Nadav, Vardi, Gal, & Shamir, Ohad. 2022. *Implicit Regularization Towards Rank Minimization in ReLU Networks*.

Vardi, Gal. 2023. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, **66**(6), 86–93.

Vardi, Gal, & Shamir, Ohad. 2021. *Implicit Regularization in ReLU Networks with the Square Loss*.

Vasudeva, Bhavya, Deora, Puneesh, & Thrampoulidis, Christos. 2025. *Implicit Bias and Fast Convergence Rates for Self-attention*.

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

Wei, William WS. 2013. *Time series analysis*. Oxford University Press.

# A Extra Discussions and Comparisons

## A.1 Concrete Examples for Models

### A.1.1 Linear Autoregressive

Let $\mathbf{x} \in \mathbb{R}^l$, $\mathbf{w} \in \mathbb{R}^D$. Suppose $D = l = 2$ and consider $T = 2$ autoregressive steps. We have that:

$$x_3 = \langle \mathbf{w}, \mathbf{x} \rangle = w_1 x_1 + w_2 x_2 \qquad x_4 = \langle \mathbf{w}, \mathbf{x}_{2:D+T-1} \rangle = \langle \mathbf{w}, \mathbf{x}_{2:3} \rangle = w_1 w_2 x_1 + (w_1 + w_2^2) x_2$$

For our example the polynomial vector is $\mathcal{P}(\mathbf{w}) = \begin{pmatrix} w_1 w_2 \\ w_1 + w_2^2 \end{pmatrix}$, and in our example $x_4 = (w_1 w_2, w_1 + w_2^2)\ \mathbf{x}$.

### A.1.2 Autoregressive SSM

For a concrete example, consider $l = 2, T = 2$. Then we have

$$x_3 = C^\top A B x_1 + C^\top B x_2 \qquad x_4 = C^\top A^2 B x_1 + C^\top A B x_2 + C^\top B x_3 = C^\top A^2 B x_1 + C^\top A B x_2 + C^\top B \left( C^\top A B x_1 + C^\top B x_2 \right) = \left( C^\top A^2 B + C^\top B C^\top A B \right) x_1 + \left( C^\top A B + \left( C^\top B \right)^2 \right) x_2$$

That is the corresponding polynomial would be

$$Q(A, B, C) = \begin{pmatrix} C^\top A^2 B + C^\top B C^\top A B \\ C^\top A B + (C^\top B)^2 \end{pmatrix}$$

## A.2 Comparison to Kunin *et al.* (2022)

Kunin *et al.* (2022) define a model to be $\Lambda$-quasi-homogeneous if:

$$f(\mathbf{x}, e^{\alpha \Lambda} \mathbf{w}) = e^{\alpha} f(\mathbf{x}, \mathbf{w})$$

For any scalar $\alpha > 0$, input $\mathbf{x}$ and parameters $\mathbf{w}$. These models satisfy a type of Euler-Homogeneous-Theorem. i.e.

$$\langle \nabla_{\mathbf{w}} f(\mathbf{x}, \mathbf{w}), \Lambda \mathbf{w} \rangle = f(\mathbf{x}, \mathbf{w})$$

Under the setting of Section 3.2, let us consider the model with two autoregressive steps, with an example in 2D, i.e. $f(\mathbf{x}, \mathbf{w}) = x_1 w_1 w_2 + x_2 (w_1 + w_2^2)$. Since $f$ is linear in $\mathbf{x}$, and the identity must hold for any $\mathbf{x}$, then the identities

$$\langle \nabla_{\mathbf{w}} w_1 w_2, \Lambda \mathbf{w} \rangle = w_1 w_2$$
$$\left\langle \nabla_{\mathbf{w}} w_1 + w_2^2, \Lambda \mathbf{w} \right\rangle = w_1 + w_2^2$$

must hold. Let $\Lambda = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. We will start with the first identity

$$\langle \nabla_{\mathbf{w}} w_1 w_2, \Lambda \mathbf{w} \rangle = c w_1^2 + (a + d) w_1 w_2 + b w_2^2 \, .$$

Hence $b = c = 0, a + d = 1$. For the second identity, we have

$$\left\langle \nabla_{\mathbf{w}} w_1 + w_2^2, \Lambda \mathbf{w} \right\rangle = a w_1 + b w_2 + 2 c w_1 w_2 + 2 d w_2^2 \, .$$

Hence $a = 1, d = 1/2, b = c = 0$. But now we have $a + d = 3/2 \neq 1$, contradiction to the first identity's requirements. Hence there exists no $\Lambda$ s.t. $f$ is quasi-homogeneous.

### A.3 Fractional Effective Degree

It is important to note that while Assumptions 1 to 3 capture a wider variety of dynamical behaviors than Cai *et al.* (2025), they still do not capture the full array of possible dynamics. For example, our effective degree assumption does not always hold.

**Lemma A.1.** *There exists an example* $(\mathbf{x}, y)$*, under the setting in Section 3.2, using Equation (1) after initializing at* $\mathbf{w}^{(0)} = (-2, 1)$*, with a constant step size* $\eta > 0$ *s.t. its margin* $m(\mathbf{w}^{(t)}) = \Theta\left(\left\|\mathbf{w}^{(t)}\right\|^{3/2}\right)$*, violating both Assumption 4 and Assumption 3.*

We take $\mathbf{x} = \begin{pmatrix} 0 \\ -1/3 \end{pmatrix}$, $y = 1$, and

$$\mathcal{P}(\mathbf{w}) = \begin{pmatrix} w_1^2 + w_1 w_2^2 \\ 2w_1 w_2 + w_2^3 \end{pmatrix}$$

corresponding to 3 autoregressive steps. We have $\mathcal{L}(\mathbf{w}) = \exp(\frac{2}{3}w_1 w_2 + \frac{1}{3}w_2^3)$. The idea is that the optimization path follows a coordinate-relation of $w_1 \sim -w_2^2$, so we have $\bar{\mathbf{w}}^\infty = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$ and $\|\mathbf{w}\| \sim w_2^2$, both since $w_1$ dominates $w_2$. Similarly, substituting $w_1$ with $-w_2^2$ we get that the margin $\sim w_2^3$. Hence effectively behaves like a $3/2$ degree polynomial - which is not an integer. A way to think about what happens is that $\mathbf{w}^{(t)}/\left\|\mathbf{w}^{(t)}\right\|$ converges slow enough, that the radial speed in the direction of $\bar{\mathbf{w}}^\infty$ is significantly slower than the radial speed of $\mathbf{w}$. The full proof is in Appendix A.3.1. Note that in this example Assumption 4 does not hold as well, as by Lemma C.1 it requires the margin to scale like $\Theta\left(\left\|\mathbf{w}^{(t)}\right\|^2\right)$.

A natural question arises as to whether Assumption 3 can be generalized to fractional-effective degrees. This is an interesting future research direction. We note that it cannot simply be extended using our proof technique. First, note that there is no trivial generalization for homogenization to a fractional degree of a polynomial with integer powers, and Assumption 3 requires the limit to equal a homogenization of the margin. Second, if we relax Assumption 4 to only require $\lim_{t\to\infty} \frac{m(\mathbf{w}^{(t)})}{\|\mathbf{w}^{(t)}\|^k} = C \neq 0$ for some fractional $k$, a new problem arises. Recall that by Assumption 2 we have that $\frac{\mathbf{w}^{(t)}}{\|\mathbf{w}^{(t)}\|} = \bar{\mathbf{w}}^\infty + \delta^{(t)}$ with $\delta^{(t)} \to 0$, then using taylor approximation we have

$$\begin{aligned}
\nabla m\left(\mathbf{w}^{(t)}\right) &= \sum_{i=1}^{d} \left\|\mathbf{w}^{(t)}\right\|^{i-1} \nabla m^i \left(\frac{\mathbf{w}^{(t)}}{\left\|\mathbf{w}^{(t)}\right\|}\right) \\
&= \sum_{i=1}^{d} \left\|\mathbf{w}^{(t)}\right\|^{i-1} \nabla m^i \left(\bar{\mathbf{w}}^\infty + \delta^{(t)}\right) \\
&= \sum_{i=1}^{d} \left\|\mathbf{w}^{(t)}\right\|^{i-1} \sum_{j=1}^{i} \nabla^j m^i \left(\bar{\mathbf{w}}^\infty\right) \left[(\delta^{(t)})^{\otimes j-1}\right]
\end{aligned}$$

Where $(\delta^{(t)})^{\otimes j} = (\delta^{(t)}, \ldots, \delta^{(t)})$ $i$ times. Now $\left\|\nabla^j m^i \left(\bar{\mathbf{w}}^\infty\right) \left[(\delta^{(t)})^{\otimes j}\right]\right\| = O\left(\left\|\delta^{(t)}\right\|^{j-1}\right)$, and hence if $\bar{\mathbf{w}}^\infty$ zeros out some of the top homogenizations of $m$ and their gradients, the dominant terms may come from either the expansions of lower degree homogenizations, or expansions of those canceled top homogenizations, i.e. $\delta^{(t)} \left\|\mathbf{w}^{(t)}\right\|^k \overset{?}{\lesssim} \left\|\mathbf{w}^{(t)}\right\|^{k-1}$ depending on how fast or slow $\delta^{(t)} \to 0$. All in all, we find that there's a wide range of possible dynamical behaviors when optimizing polynomials with the exponential loss, some of which change the resulting implicit bias as we see in Section 5. We leave it for future work to investigate these cases, and focus on Assumption 3. We verify all of our assumptions experimentally in Section 6.

### A.3.1 Proof of Lemma A.1

In the following example, we demonstrate that in a simple setting, our effective degree assumption may fail. Our example is in $2D$ and 3 linear autoregressive steps. We take $\mathbf{x} = \begin{pmatrix} 0 \\ -1/3 \end{pmatrix}$, $y = 1$, and

$$\mathcal{P}(\mathbf{w}) = \begin{pmatrix} w_1^2 + w_1 w_2^2 \\ 2w_1 w_2 + w_2^3 \end{pmatrix}$$

So we have $\mathcal{L}(\mathbf{w}) = \exp(\frac{2}{3} w_1 w_2 + \frac{1}{3} w_2^3)$. The idea is that the optimization path follows a relation of $w_1 \sim -w_2^2$, so we have $\bar{\mathbf{w}}^\infty = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$ and $\|\mathbf{w}\| \sim w_2^2$ while the margin $\sim w_2^3$, hence the effective degree is $3/2$, not an integer. A way to think about what happens is that $\mathbf{w}^{(t)} / \left\|\mathbf{w}^{(t)}\right\|$ converges slow enough, that the radial speed in the direction of $\bar{\mathbf{w}}^\infty$ is significantly slower than the radial speed of $\mathbf{w}$. Let us analyze this formally:

$$\frac{\partial \mathcal{L}}{\partial w_1} = \frac{2}{3} \cdot \mathcal{L}(\mathbf{w}) w_2$$
$$\frac{\partial \mathcal{L}}{\partial w_2} = \mathcal{L}(\mathbf{w}) \left( \frac{2}{3} w_1 + w_2^2 \right)$$

For a fixed learning rate $\eta$, denote $\alpha^{(t)} = \eta \mathcal{L}(\mathbf{w}^{(t)})$, then the gradient descent update gives

$$w_1^{(t+1)} = w_1^{(t)} - \alpha^{(t)} \frac{2}{3} w_2$$
$$w_2^{(t+1)} = w_2^{(t)} - \alpha^{(t)} \left( \frac{2}{3} w_1 + w_2^2 \right)$$

Let us start the optimization from $\mathbf{w} = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$. Note that if $d := -\frac{2}{3} w_1 - w_2^2 \geq 0$ and $w_2 \geq 0$, then $w_2$ is non-decreasing. We will show $d^{(t)} \geq 0 \; \forall t$. Clearly $d^{(0)} = 1/3$. Let us see how $d$ evolves:

$$\begin{aligned} d^{(t+1)} &= -\frac{2}{3} w_1^{(t+1)} - \left( w_2^{(t+1)} \right)^2 \\ &= -\frac{2}{3} w_1^{(t)} + \alpha^{(t)} \frac{4}{9} w_2^{(t)} - \left( w_2^{(t)} + \alpha^{(t)} d^{(t)} \right)^2 \\ &= -\frac{2}{3} w_1^{(t)} + \alpha^{(t)} \frac{4}{9} w_2^{(t)} - \left( w_2^{(t)} \right)^2 - 2 w_2^{(t)} \alpha^{(t)} d^{(t)} - \left( \alpha^{(t)} d^{(t)} \right)^2 \\ &= d^{(t)} - \left( \alpha^{(t)} d^{(t)} \right)^2 + \alpha^{(t)} w_2^{(t)} \left( \frac{4}{9} - 2 d^{(t)} \right) \end{aligned} \tag{4}$$

So whenever $d^{(t)} = 0$, $d^{(t+1)} = \frac{4}{9} \alpha^{(t)} w_2^{(t)} > 0$ as long as $w_2^{(t)}$ is positive. First we will show that $d^{(t)}$ is upper bounded. If $d^{(t)} \geq 4/9$, then $4/9 - 2d^{(t)} \leq -d^{(t)}$ and so by Equation (4), $d^{(t+1)} \leq d^{(t)} - \alpha^{(t)} w_2^{(t)} d^{(t)} \leq d^{(t)}$ as long as $w_2^{(t)}$ is positive. Hence $d^{(t)} \leq \max\{4/9, d^{(0)}\} = 4/9$.

Now whenever $d^{(t)} \geq 0$, we have

$$d^{(t+1)} \geq d^{(t)} - \left( \alpha^{(t)} d^{(t)} \right)^2 - 2 \alpha^{(t)} d^{(t)} w_2^{(t)} = d^{(t)} \left( 1 - d^{(t)} \left( \alpha^{(t)} \right)^2 - 2 \alpha^{(t)} w_2^{(t)} \right)$$

so it's enough to show $-d^{(t)} \left( \alpha^{(t)} \right)^2 - 2 \alpha^{(t)} w_2^{(t)} \geq -1$ for $d^{(t)}$ to stay non-negative. Note that we can rewrite $\frac{2}{3} w_1^{(t)} w_2^{(t)} + \frac{1}{3} \left( w_2^{(t)} \right)^3 = -\frac{2}{3} \left( w_2^{(t)} \right)^3 - d^{(t)} w_2^{(t)}$. Let us bound

$$\begin{aligned} \alpha^{(t)} w_2^{(t)} &\leq \eta w_2^{(t)} \exp\left( \frac{2}{3} w_1^{(t)} w_2^{(t)} + \frac{1}{3} \left( w_2^{(t)} \right)^3 \right) \\ &= \eta w_2^{(t)} \exp\left( -\frac{2}{3} \left( w_2^{(t)} \right)^3 - d^{(t)} w_2^{(t)} \right) \\ &\leq \eta w_2^{(t)} \exp\left( -\frac{2}{3} \left( w_2^{(t)} \right)^3 \right) \leq 0.6\eta \end{aligned} \tag{5}$$

Where the last inequality is by finding the maximum of $w_2^{(t)} \exp\left(-\frac{2}{3}\left(w_2^{(t)}\right)^3\right)$ for $w_2^{(t)} \geq 0$. Hence choosing for example $\eta = 0.1$ we get

$$-d^{(t)}\left(\alpha^{(t)}\right)^2 - 2\alpha^{(t)} w_2^{(t)} \geq -4/9\eta^2 - 2\alpha^{(t)} w_2^{(t)} \geq -0.1 - 0.2 \geq -1$$

So we have that $d^{(t)} \geq 0$, and so $w_2^{(t)}$ is non-decreasing. We want to show that it's unbounded. Assume towards contradiction that it is bounded, i.e. $w_2^{(t)} \leq U$, then we have that $\mathcal{L}(\mathbf{w}^{(t)}) = \exp\left(-\frac{2}{3}\left(w_2^{(t)}\right)^3 - d^{(t)} w_2^{(t)}\right) \geq \mathcal{L}_{\inf} > 0$. Now

$$w_1^{(t)} - \frac{2}{3}\alpha^{(t)} w_2^{(t)} \leq w_1^{(t)} - \frac{2}{3}\alpha^{(t)}$$

Since we start from $w_2^{(0)} = 1$ and it's non-decreasing. Finally, we get

$$w_1^{(t+1)} \leq w_1^{(t)} - \frac{2}{3}\eta\mathcal{L}_{\inf}$$

So $w_1^{(t)} \to -\infty$, and hence $d^{(t)} = -\frac{2}{3}w_1^{(t)} - \left(w_2^{(t)}\right)^2 \to \infty$, contradicting its upper bound. Hence we have $w_2^{(t)} \to \infty$, and similarly, $\mathcal{L}(\mathbf{w}^{(t)}) \to 0$.

Now consider the following sequence

$$e^{(t)} = d^{(t)} - 2/9$$

We want to show it converges to 0. Note that

$$e^{(t+1)} - e^{(t)} = -\left(\alpha^{(t)} d^{(t)}\right)^2 + \alpha^{(t)} w_2^{(t)}\left(\frac{4}{9} - 2d^{(t)}\right) = -\left(\alpha^{(t)} d^{(t)}\right)^2 - 2\alpha^{(t)} w_2^{(t)} e^{(t)}$$

Hence

$$e^{(t+1)} = e^{(t)}\left(1 - 2\alpha^{(t)} w_2^{(t)}\right) - \left(\alpha^{(t)} d^{(t)}\right)^2$$

Now from Equation (5), and positivity of $w_2^{(t)}$ we have that for large enough $t$, $\left(1 - 2\alpha^{(t)} w_2^{(t)}\right) \in (0, 1)$. We want to show that $\sum_i \left(\alpha^{(t)} d^{(t)}\right)^2 < \infty$. Recall that

$$w_2^{(t+1)} - w_2^{(t)} = \alpha^{(t)} d^{(t)} \leq \frac{4}{9}\eta \exp\left(-\frac{2}{3}\left(w_2^{(t)}\right)^3\right)$$

Let $I_n := \{t : w_2^{(t)} \in [n, n+1)\}$, then for large enough $n$, we have that $\sum_{t\in I_n}\left(w_2^{(t+1)} - w_2^{(t)}\right) = \sum_{t\in I_n} \alpha^{(t)} d^{(t)} \leq 2$. Moreover, $\max_{t\in I_n} \alpha^{(t)} d^{(t)} \leq \frac{4}{9}\eta \exp\left(-\frac{2}{3}n^3\right)$ so overall

$$\sum_{t\in I_n}\left(\alpha^{(t)} d^{(t)}\right)^2 \leq \max_{t\in I_n} \alpha^{(t)} d^{(t)} \sum_{t\in I_n}\left(\alpha^{(t)} d^{(t)}\right) \leq \frac{8}{9}\eta \exp\left(-\frac{2}{3}n^3\right)$$

and overall

$$\sum_t \left(\alpha^{(t)} d^{(t)}\right)^2 = \sum_n \sum_{t\in I_n}\left(\alpha^{(t)} d^{(t)}\right)^2 \leq \sum_n \frac{8}{9}\eta \exp\left(-\frac{2}{3}n^3\right) < \infty$$

Note that since $-\left(w_2^{(t)}\right)^2 - \frac{2}{3}w_1^{(t)} = d^{(t)} \geq 0$ we have that $w_1^{(t)} \to -\infty$. And

$$w_1^{(t)} = w_1^{(0)} - \frac{2}{3}\sum_{k=0}^{t-1} \alpha^{(k)} w_2^{(k)} \to -\infty$$

So overall $\sum_{k=0}^{t-1} \alpha^{(k)} w_2^{(k)} \to \infty$. Now let us define $e^{(t)} := d^{(t)} - 2/9$, we want to use Lemma F.4 to show $e^{(t)} \to 0$. We have

$$e^{(t+1)} = e^{(t)}(1 - 2\alpha^{(t)} w_2^{(t)}) - \left(\alpha^{(t)} d^{(t)}\right)^2$$

so indeed using Lemma F.4 we get $d^{(t)} \to 2/9$. But now

$$w_1^{(t)} = -\frac{3}{2}\left(w_2^{(t)}\right)^2 - 1/3 + o(1)$$

and so

$$\frac{2}{3}\left(w_2^{(t)}\right)^3 + d^{(t)} w_2^{(t)} = \Theta\left(\left(w_2^{(t)}\right)^3\right)$$

but $\left\|\mathbf{w}^{(t)}\right\| = \Theta\left(\left(w_2^{(t)}\right)^2\right)$, so the margin is $\Theta\left(\left\|\mathbf{w}^{(t)}\right\|^{3/2}\right)$

### A.4 Proof of Lemma 4.4

Lemma C.1 implies Assumption 1. Theorem 6.2 in Cai *et al.* (2025) (Theorem F.2) implies Assumption 2.

**For the case where $q$ is $M$-homogeneous** these two imply Assumption 3 immediately by homogeneity.

**For the case where $q$ is not homogeneous and** $\deg(p) \geq 1$ let us bound the loss explicitly:

$$\begin{aligned}\mathcal{L}(\mathbf{w}^{(t)}) &= \sum_{i=1}^{N} \exp\left(-y_i \left\langle q(\mathbf{w}^{(t)}), \mathbf{x}_i \right\rangle\right) \\ &\geq \max_{i\in[N]} \exp\left(-y_i \left\langle q(\mathbf{w}^{(t)}), \mathbf{x}_i \right\rangle\right) \end{aligned} \tag{6}$$

We have the margin as defined in Cai *et al.* (2025) is,

$$\gamma(t) := \frac{\exp\left(\Phi\left(\exp\left(p\left(\left\|\mathbf{w}^{(t)}\right\|\right)\right)\mathcal{L}\left(\mathbf{w}^{(t)}\right)/N\right)\right)}{\left\|\mathbf{w}^{(t)}\right\|^M}$$

Where $M = \deg(q)$, and $\Phi(x) = \log(\log(1/Nx)) + 2/\log(Nx)$. Since $\exp\circ\Phi(x)$ is monotonous decreasing, using Equation (6) and for large enough $t$ s.t. $p\left(\left\|\mathbf{w}^{(t)}\right\|\right) \geq 1$,

$$\begin{aligned}\gamma(t) &\leq \frac{\exp\left(\Phi\left(\max_{i\in[N]}\exp\left(-y_i\left\langle q(\mathbf{w}^{(t)}), \mathbf{x}_i\right\rangle\right)/N\right)\right)}{\left\|\mathbf{w}^{(t)}\right\|^M} \\ &= \frac{\exp\left(\log\log\left(\frac{1}{\max_{i\in[N]}\exp\left(-y_i\left\langle q(\mathbf{w}^{(t)}),\mathbf{x}_i\right\rangle\right)}\right) + \frac{2}{\log\left(\max_{i\in[N]}\exp\left(-y_i\left\langle q(\mathbf{w}^{(t)}),\mathbf{x}_i\right\rangle\right)\right)}\right)}{\left\|\mathbf{w}^{(t)}\right\|^M} \\ &= \frac{\exp\left(\log\left(\min_{i\in[N]} y_i\left\langle q(\mathbf{w}^{(t)}), \mathbf{x}_i\right\rangle\right) + \frac{2}{\max_{i\in[N]}\left(-y_i\left\langle q(\mathbf{w}^{(t)}),\mathbf{x}_i\right\rangle\right)}\right)}{\left\|\mathbf{w}^{(t)}\right\|^M} \\ &\leq \frac{\exp\left(\log\left(\min_{i\in[N]} y_i\left\langle q(\mathbf{w}^{(t)}), \mathbf{x}_i\right\rangle\right)\right)}{\left\|\mathbf{w}^{(t)}\right\|^M} \qquad (7) \\ &= \frac{\min_{i\in[N]} y_i\left\langle q(\mathbf{w}^{(t)}), \mathbf{x}_i\right\rangle}{\left\|\mathbf{w}^{(t)}\right\|^M} \qquad (8)\end{aligned}$$

Where Equation (7) is due to $\frac{2}{\max_{i\in[N]}\left(-y_i\left\langle q(\mathbf{w}^{(t)}),\mathbf{x}_i\right\rangle\right)}$ being negative for large enough $t$. Taking the limit on both sides of Equation (8), and using Theorem F.1 showing that $\gamma(t)$ is positive and increasing for large enough $t$, and since we have that $\deg(q) = M$, then the normalized margin converges to some positive constant. For simplicity let us denote the margin polynomials $m_i(\mathbf{w}) := y_i\left\langle q(\mathbf{w}^{(t)}), \mathbf{x}_i\right\rangle$, then using homogeneity we have

$$\frac{m(\mathbf{w}^{(t)})}{\left\|\mathbf{w}^{(t)}\right\|^M} = \sum_{i=0}^{M}\left\|\mathbf{w}^{(t)}\right\|^{i-M} m^i\left(\frac{\mathbf{w}^{(t)}}{\left\|\mathbf{w}^{(t)}\right\|}\right)$$

Taking the limit, one summand must not go to zero (because we know the limit is positive), hence it must be the last one, and we get $\lim_{t\to\infty} \frac{m(\mathbf{w}^{(t)})}{\left\|\mathbf{w}^{(t)}\right\|^M} = m^M(\bar{\mathbf{w}}^\infty) > 0$. Similarly, from homogeneity we get that $\left\langle \bar{\mathbf{w}}^\infty, \nabla m^M(\bar{\mathbf{w}}^\infty) \right\rangle = M m^M(\bar{\mathbf{w}}^\infty) > 0$, hence $\nabla m^{M-1}(\bar{\mathbf{w}}^\infty) \neq 0$, and must dominate the gradient. Thus, we have concluded Assumption 3.

**For the case where $q$ is not homogeneous and $\deg(p) = 0$** We have that $p\left(\left\|\mathbf{w}^{(t)}\right\|\right)$ is a constant and hence

$$\gamma(t) := \frac{\exp\left(\Phi\left(C\mathcal{L}\left(\mathbf{w}^{(t)}\right)/N\right)\right)}{\left\|\mathbf{w}^{(t)}\right\|^M}$$

For some $C > 0$. The rest of the argument follows exactly as above, with a trailing constant $C$, that does not change the sign of the limits, leading to the same conclusion.

# B Multidimensional Tokens

## B.1 Linear Autoregressive Models

Given an example $X \in \mathbb{R}^{l\times D}$ which we interpret as the initial part of some sequence $X = X_{[1:D]} = (\mathbf{x}_1, \ldots, \mathbf{x}_D)$ (i.e $\mathbf{x}_i$ denote the $i^{th}$ column of $X$), we compute a sequence of $T$ autoregressive steps, the final's sign being our label. Formally, the first step will be $\mathbf{x}_{D+1} = X\mathbf{w}$. The second step $X_{D+2} = X_{2:D+1}\mathbf{w}$, and at the $T^{th}$ step we'll have $x_{D+T} = X_{T:D+T-1}\mathbf{w}$. For example, let $D = 2$ and let us consider two autoregressive steps ($T = 2$). We have that

$$\begin{aligned}
\mathbf{x}_3 &= X\mathbf{w} = \begin{pmatrix} w_1x_{1,1} + w_2x_{1,2} \\ w_1x_{2,1} + w_2x_{2,2} \end{pmatrix} \\
\mathbf{x}_4 &= X_{[2:D+T-1]}\mathbf{w} = X_{[2:3]}\mathbf{w} \\
&= \begin{pmatrix} w_1x_{1,2} + w_2x_{1,3} \\ w_1x_{2,2} + w_2x_{2,3} \end{pmatrix} = \\
&= \begin{pmatrix} w_1w_2x_{1,1} + (w_1 + w_2^2)x_{1,2} \\ w_1w_2x_{2,1} + (w_1 + w_2^2)x_{2,2} \end{pmatrix}
\end{aligned}$$

Note that in contrast to Section 3.2, the resulting $\mathbf{x}_4$ cannot serve as a label by itself, and hence we must fix, or learn a projection vector $\mathbf{p} \in \mathbb{R}^l$, s.t. $y = \operatorname{sgn}(\langle \mathbf{p}, \mathbf{x}_4 \rangle)$. We denote the vector of coefficients of $x_1, \ldots, x_D$ in $x_{D+T}$ as $\mathcal{P}$, a polynomial in $\mathbf{w}$, and optionally $\mathbf{p}$. For our running example that vector would be $\mathcal{P}(\mathbf{w}, \mathbf{p}) = \begin{pmatrix} w_1w_2p_1 & w_1p_1 + w_2^2p_1 \\ w_1w_2p_2 & w_1p_2 + w_2^2p_2 \end{pmatrix}$. Using this notation, we have that $\langle \mathcal{P}(\mathbf{w}, \mathbf{p}), X \rangle = x_{D+T}$, and in our example $x_4 = \operatorname{Tr}\left(\begin{pmatrix} w_1w_2p_1 & w_1p_1 + w_2^2p_1 \\ w_1w_2p_2 & w_1p_2 + w_2^2p_2 \end{pmatrix}\begin{pmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \end{pmatrix}^\top\right)$. Note that for a dataset $\{X_i, y_i\}_{n=1}^N$ we denote the polynomial corresponding to the $i^{th}$-example as $\mathcal{P}_i$, but in this autoregressive setting, it is the case that $\mathcal{P}_i = \mathcal{P}_j$ for $i \neq j$, if both examples $i, j$ are labeled by the same number $T$ of autoregressive steps. We state the theorem for the case of learning $\mathbf{p}$, but it can easily be amended for fixed $\mathbf{p}$.

**Theorem B.1** (Implicit Bias of Multidimensional Linear Autoregressive Models)**.** *For parameters $\mathbf{w} \in \mathbb{R}^D$ and $\mathbf{p} \in \mathbb{R}^l$ denote $\theta := (\mathbf{w}, \mathbf{p})$, let $\{X_i, y_i\}_{i=1}^N$, where $y_i = \operatorname{sgn}\langle \mathcal{P}_i(\theta), X_i \rangle$. Consider the sequence of gradient updates according to Equation (1), with any non-increasing sequence of step sizes $\{\eta_t\}$, minimizing the empirical risk. If Assumptions 1 to 3, then $\bar{\theta}^\infty := \lim_{t\to\infty} \frac{\theta^{(t)}}{\left\|\theta^{(t)}\right\|}$ is a positive scaling of a first order stationary point of the following optimization problem:*

$$\begin{aligned}
\min_{\theta\in\mathbb{R}^{D+l}} \|\theta\|_2^2 \ \text{ s.t. } & \forall i, y_i \left\langle X_i, \mathcal{P}_i^{y_i}(\theta) \right\rangle \geq 0 \\
& \forall i \in Z, y_i C_i \left\langle X_i, \mathcal{P}_i^{y_i}(\theta) \right\rangle \geq 1
\end{aligned}$$

*where $Z = \arg\min_{i\in[n]} \left\{\operatorname{effdeg}_\theta\left(\langle \mathcal{P}_i, \mathbf{x}_i y_i \rangle\right)\right\}$.*

*Proof.* The proof follows immediately from Theorem 5.1, by concatenating $\theta = (\mathbf{w}, \mathbf{p})$ into one vector, concatenating each $X_i$'s columns into a vector $\mathbf{v}_i$, and concatenating $\mathcal{P}_i$'s columns into a vector $\tilde{\mathcal{P}}_i$ the same way, such that they match and then $\left\langle \mathbf{v}_i, \tilde{\mathcal{P}}_i(\theta) \right\rangle = \langle X_i, \mathcal{P}_i(\mathbf{w}, \mathbf{p}) \rangle$. □

### B.2 Autoregressive SSMs

Here we consider the full Autoregressive SSMs, as it subsumes the simpler SSM case. We repeat the notation for convenience. We denote the parameters of the model by $A \in \mathbb{R}^{d_A \times d_A}$, $B \in \mathbb{R}^{d_A \times d}$, $C \in \mathbb{R}^{d_A \times d}$. Consider an example $X \in \mathbb{R}^{d \times l}$, that we can think of as an input prompt, or as a beginning of some sequence $X = X_{1:l} = \begin{pmatrix} x_{1,1} & \dots & x_{1,l} \\ \vdots & \ddots & \vdots \\ x_{d,1} & \dots & x_{d,l} \end{pmatrix}$, we compute a sequence of $T$ autoregressive steps:

$$(X)_{l+1} = C^\top \sum_{i=1}^{l} A^{l-i} B(X)_i, \quad (X)_{l+2} = C^\top \sum_{i=1}^{l+1} A^{l+1-i} B(X)_i, \quad \dots \quad , (X)_{l+T} = C^\top \sum_{i=1}^{l+T-1} A^{l+T-1-i} B(X)_i \,. \tag{9}$$

Where $(X)_i$ denotes the $i^{th}$ column of $X$ The label of the sequence $\mathbf{x}$ is defined to be $\operatorname{sgn}(\langle \mathbf{x}_{l+T}, \mathbf{p} \rangle)$, where $\mathcal{P}$ is some projection vector that is required in order to get a scalar prediction. This vector may be learned or fixed. We phrase the theorem for the general case, where $\mathcal{P}$ can be learned

**Theorem B.2** (Implicit bias of SSMs)**.** *For parameters $A \in \mathbb{R}^{d_A \times d_A}$, $B \in \mathbb{R}^{d_A \times d}$, $C \in \mathbb{R}^{d_A \times d}$ and $\mathbf{p} \in \mathbb{R}^d$ denoted for simplicity by $\mathbf{w} := (A, B, C, \mathbf{p})$, let $\{X_i, y_i\}_{i=1}^N$, where $y_i = \operatorname{sgn} \langle Q_i(\mathbf{w}), X_i \rangle$. Consider the sequence of gradient updates according to Equation (1), with a non-increasing sequence of step sizes $\{\eta_t\}$, minimizing the empirical risk. If Assumptions 1 to 3, $\bar{\mathbf{w}}^\infty := \lim_{t \to \infty} \frac{\mathbf{w}^{(t)}}{\|\mathbf{w}^{(t)}\|}$ is a positive scaling of a first order stationary point of the following optimization problem:*

$$\min_{\mathbf{w} \in \mathbb{R}^{d_A \times d_A + 2d_A \times d + d}} \|\mathbf{w}\|_2^2 \ \text{s.t.} \ \forall i, y_i \left\langle X_i, Q_i^{\nu_i}(\mathbf{w}) \right\rangle \geq 0$$
$$\forall i \in Z, y_i C_i \left\langle X_i, Q_i^{\nu_i}(\mathbf{w}) \right\rangle \geq 1$$

*where $Z = \arg\min_{i \in [n]} \left\{ \operatorname{effdeg}_{\mathbf{w}} \left( \langle Q_i, \mathbf{x}_i \rangle \right) \right\}$.*

*Proof.* The proof follows immediately from Theorem 5.1, by concatenating $\mathbf{w} = (A, B, C, \mathbf{p})$ into one vector, and noting that since at each step we only multiply by the parameters and not by $X_i$, we can write a coefficient vector for every $x_{i,j}$ in terms of $A, B, C$ and $\mathbf{p}$ only, then using that vector as our polynomial, and concatenating each $X_i$ to a long vector respecting these coefficients. □

## C Proof of Lemma 4.3

Consider the following simple toy example in 4D: Take one example $\mathbf{x} = (0, 0, 0, 1)$, $y = 1$ with 4-autoregressive steps. We have that

$$\begin{aligned}
x_5 &= w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 \\
x_6 &= w_1 x_2 + w_2 x_3 + w_3 x_4 + w_4 x_5 = w_3 + w_4^2 \\
x_7 &= w_1 x_3 + w_2 x_4 + w_3 x_5 + w_4 x_6 = w_2 + 2 w_3 w_4 + w_4^3 \\
x_8 &= w_1 x_4 + w_2 x_5 + w_3 x_6 + w_4 x_7 = w_1 + 2 w_2 w_4 + w_3^2 + 3 w_3 w_4^2 + w_4^4
\end{aligned}$$

So we have that

$$\mathcal{L}(\mathbf{w}) = \exp(-w_1 - 2 w_2 w_4 - w_3^2 - 3 w_3 w_4^2 - w_4^4)$$

The idea behind this toy example is that

$$-\nabla \mathcal{L}(\mathbf{w}) = \mathcal{L}(\mathbf{w}) \cdot \begin{pmatrix} 1 \\ 2 w_4 \\ 2 w_3 + 3 w_4^2 \\ 2 w_2 + 6 w_3 w_4 + 4 w_4^3 \end{pmatrix}$$

Hence whenever $w_2 = w_3 = w_4 = 0$ we get that

$$-\nabla\mathcal{L}(\mathbf{w}) = \mathcal{L}(\mathbf{w}) \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

That is, $w_2, w_3, w_4$ remain 0. This produces the dynamic that while $\deg(\mathcal{P}(\mathbf{w})^\top \mathbf{x}y) = 4$, we have that $\operatorname{effdeg}_{\mathbf{w}}(\mathcal{P}(\mathbf{w})^\top \mathbf{x}y) = 1$. Effectively the higher parts are not used during optimization.

We will first show why the assumptions of Cai *et al.* (2025) fail, and proceed to show why our assumptions hold.

### C.1 Proof of Lemma 4.3

Let us first restate the assumption:

**Assumption 5** (Cai *et al.* (2025), simplified for our example)**.** *At some point $t_0$ during optimization, $\mathcal{L}(\mathbf{w}^{(t_0)}) \leq \exp\left(-p\left(\left\|\mathbf{w}^{(t_0)}\right\|\right)\right)$ for some $p(x)$. If the predictor is a polynomial of the form $f(\mathbf{w}; \mathbf{x}) = \langle q(\mathbf{w}), \mathbf{x}\rangle$, then $\deg(p) = \deg(q) - 1$.*

They later use this assumption to prove the following:

**Lemma C.1** (Cai *et al.* (2025), simplified)**.** *$\forall t \geq t_0$ we have that $\mathcal{L}(\mathbf{w}^{(t)}) \leq \exp\left(-p\left(\left\|\mathbf{w}^{(t)}\right\|\right)\right)$*

Let us analyze the example in light of Assumption 4. First, recall that

$$\mathcal{L}(\mathbf{w}) = \exp(-w_1 - 2w_2w_4 - w_3^2 - 3w_3w_4^2 - w_4^4)$$

and

$$-\nabla\mathcal{L}(\mathbf{w}) = \mathcal{L}(\mathbf{w}) \cdot \begin{pmatrix} 1 \\ 2w_4 \\ 2w_3 + 3w_4^2 \\ 2w_2 + 6w_3w_4 + 4w_4^3 \end{pmatrix}$$

Initializing at $\mathbf{w}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ and whenever $w_2 = w_3 = w_4 = 0$, we get

$$-\nabla\mathcal{L}(\mathbf{w}) = \mathcal{L}(\mathbf{w}) \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{10}$$

So assuming a constant step size $\eta$,

$$\begin{aligned} w_1^{(t+1)} &= w_1^{(t)} + \eta\mathcal{L}(\mathbf{w}^{(t)}) \\ w_j^{(t+1)} &= w_j^{(t)} + 0 \quad j \in \{2, 3, 4\} \end{aligned} \tag{11}$$

Where the updates follow from inductively $w_2^{(t)} \equiv w_3^{(t)} \equiv w_4^{(t)} \equiv 0$ staying zero for any $t$. This means that during training we have:

$$\left\|\mathbf{w}^{(t)}\right\| = \left|w_1^{(t)}\right| \tag{12}$$

Furthermore, $w_1^{(t)}$ is strictly increasing. We claim that $-w_1^{(t)} \to -\infty$. Assume otherwise, then $w_1^{(t)} \to L \in \mathbb{R}$, but then the steps can be bounded from below i.e. $\mathbf{w}_1^{(t)} - \mathbf{w}_1^{(t-1)} \geq \eta \exp^{-L} > 0$, which is a contradiction to $w_1^{(t)} \to L$. So we have

$$-w_1^{(t)} \to -\infty \tag{13}$$

But then $\mathcal{L}(\mathbf{w}^{(t)}) = \exp(-w_1^{(t)}) = \exp\left(-\left\|\mathbf{w}^{(t)}\right\|\right)$. Now assume towards contradiction that Assumption 4 holds, then by Lemma C.1 we get that there exists some $t_0$ s.t. $\forall t \geq t_0$, $\mathcal{L}(\mathbf{w}^{(t)}) \leq \exp\left(-p(\|\mathbf{w}\|)\right)$, but $\mathcal{L}(\mathbf{w}^{(t)}) = \exp\left(-\left\|\mathbf{w}^{(t)}\right\|\right) \geq \exp(-p(\|\mathbf{w}\|)$ contradiction, since $\left\|\mathbf{w}^{(t)}\right\| \to \infty$ and so for any $p$ s.t. $\deg(p) = 3$ with positive leading coefficient, we have $\left\|\mathbf{w}^{(t)}\right\| = o\left(p(\|\mathbf{w}\|)\right)$. Consequently Assumption 4 must not hold.

### C.2 Our Assumptions Hold

We want to show Assumptions 1 to 3 hold. By Equation (11) and Equation (10), we have that Assumption 2 holds, since $\nabla\mathcal{L}(\mathbf{w})$ always points in the same direction. From Equation (13) and Equation (11) we have that $\mathcal{L}(\mathbf{w}^{(t)}) = \exp(-w_1^{(t)}) \to 0$ so Assumption 1 holds. To show Assumption 3 holds true, note that

$$\lim_{t\to\infty} \frac{w_1^{(t)} + 2w_2^{(t)}w_4^{(t)} + (w_3^{(t)})^2 + 3w_3^{(t)}(w_4^{(t)})^2 + (w_4^{(t)})^4}{\left\|\mathbf{w}^{(t)}\right\|^2} = \frac{w_1^{(t)}}{(w_1^{(t)})^2} \to 0$$

On the other hand:

$$\lim_{t\to\infty} \frac{w_1^{(t)} + 2w_2^{(t)}w_4^{(t)} + (w_3^{(t)})^2 + 3w_3^{(t)}(w_4^{(t)})^2 + (w_4^{(t)})^4}{\left\|\mathbf{w}^{(t)}\right\|} = \frac{w_1^{(t)}}{\left|w_1^{(t)}\right|} = 1 = \mathcal{P}^1(\bar{\mathbf{w}}^\infty)^\top \mathbf{x}y$$

Now for the corresponding gradient,

$$\lim_{t\to\infty} \frac{\begin{pmatrix}1\\0\\0\\0\end{pmatrix}}{\left\|\mathbf{w}^{(t)}\right\|^0} = \begin{pmatrix}1\\0\\0\\0\end{pmatrix} = \nabla\mathcal{P}^1(\bar{\mathbf{w}}^\infty)^\top \mathbf{x}y$$

And hence Assumptions 1 to 3 all hold.

### C.3 Constructing General Negative Examples to Assumption 4

We can use the principles from the proof techniques of Lemma 4.3, i.e. Appendix C, to create polynomials for which similar dynamics hold. For simplicity, assume we have one example $\mathbf{x} = \begin{pmatrix}1\\1\end{pmatrix}, y = 1$. For every even $d \geq 2$ and $k < d$ we construct a polynomial such that the margin would be with degree $d$ and effective degree $k$: Take $\mathcal{P}(\mathbf{w}) = \begin{pmatrix}-(\sum_{i\in A} w_i)^k \\ (\sum_{i\in B} w_i)^d\end{pmatrix}$, for some $A \cap B = \emptyset$. The exponential loss would be:

$$\mathcal{L}(\mathbf{w}) = \exp\left(-\left(\sum_{i\in A} w_i\right)^k + \left(\sum_{i\in B} w_i\right)^d\right).$$

This will yield a similar type of dynamics. In fact, this will yield a stronger dynamic, since this will have effective degree $k$ for almost all initializations in the case of even $k$, and a constant portion of initializations in the case of odd $k$.

**Lemma C.2** (Assumption 4 does not hold for almost all initializations)**.** *Let $d \geq 2$ be an even integer, and $0 < k < d$ another even integer. Let $\mathbf{x} = \begin{pmatrix}1\\1\end{pmatrix}, y = 1$, and consider $\mathcal{P}(\mathbf{w}) = \begin{pmatrix}-(\sum_{i\in A} w_i)^k \\ (\sum_{i\in B} w_i)^d\end{pmatrix}$, for some $A \cap B = \emptyset$. Then for almost all initializations, optimizing*

$$\mathcal{L}(\mathbf{w}) = \exp\left(-\left(\sum_{i\in A} w_i\right)^k + \left(\sum_{i\in B} w_i\right)^d\right).$$

*With gradient descent yields $\mathcal{L}(\mathbf{w}^{(t)}) = \exp\left(-\Theta\left(\left\|\mathbf{w}^{(t)}\right\|^k\right)\right)$*

This can be further generalized to larger families of polynomials, building on the same insight. Importantly, our main result about implicit bias for non-homogeneous polynomials applies to all of these constructions Theorem 5.1.

*Proof.* First note that the gradient is of rank-2. Denote $u(\mathbf{w}) = \sum_{i\in A} w_i$, $v(\mathbf{w}) = \sum_{i\in B} w_i$. Then for every $i \in A$ we have

$$\frac{\partial \mathcal{L}}{\partial w_i}(\mathbf{w}) = \mathcal{L}(\mathbf{w})\frac{\partial(-u^k + v^d)}{\partial w_i}(\mathbf{w}) = -\mathcal{L}(\mathbf{w})ku^{k-1}$$

Similarly for $i \in B$, we have

$$\frac{\partial \mathcal{L}}{\partial w_i}(\mathbf{w}) = \mathcal{L}(\mathbf{w})\frac{\partial(-u^k + v^d)}{\partial w_i}(\mathbf{w}) = \mathcal{L}(\mathbf{w})du^{d-1}$$

Whenever $i \notin A \cup B$,

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0$$

Hence $\forall i, i' \in A$ $w_i - w_{i'}$ is constant. Similarly, $\forall i, i' \in B$, $w_i - w_{i'}$ is constant.

So we can re-write the update per $u, v$, i.e.

$$\begin{aligned} u_{t+1} &= u_t + \eta|A|ku_t^{k-1}\exp(-u_t^k + v^d) \\ v_{t+1} &= u_t + \eta|B|dv_t^{d-1}\exp(-u_t^k + v^d) \end{aligned} \tag{14}$$

The rest of $\mathbf{w}$ is fixed. Hence optimizing $u$ and $v$ will yield similar dynamics to optimizing $\mathbf{w}$. Denote $R := |v_0| + 1$, $M := \max\{0, -u_0^k\}$, and choose $\eta \leq \frac{\exp(-R^d - M)}{|B|dR^{d-2}}$.

Now, since $u_{t+1} = u_t + \eta|A|ku_t^{k-1}\exp(-u_t^k + v^d)$, we have that since $k$ is even, $u_t^k \geq 0$ and so $-u_t^k \leq 0 \leq M$ hence $\exp(-u_t^k + v_t^d) \leq \exp(M + v_t^d)$.

Now,

$$|v_{t+1}| = |v_t|\left|1 - \eta d\,|v_t|^{d-2}\,|B|\exp(-u_t^k - v_t^d)\right|$$

Assume $|v_t| \leq R$, then $\eta d\,|v_t|^{d-2}\,|B|\exp(-u_t^k - v_t^d) \leq 1$ from our choice of $\eta$., hence $\left|1 - \eta d\,|v_t|^{d-2}\,|B|\exp(-u_t^k - v_t^d)\right| \in [0, 1]$. So $|v_{t+1}| \leq |v_t| \leq R$, hence by induction $|v_t| \leq R$ for all $t$.

Now since $k$ is even, $k-1$ is odd and from Equation (14) we have that $u_t$ never changes sign along training and $|u_t|$ is strictly increasing. Now assume toward contradiction that $|u_t|$ is bounded, then we can bound its update by $\left|\eta|A|ku_0^{k-1}\exp(-B - R^d)\right| \leq \left|\eta|A|ku_t^{k-1}\exp(-u_t^k + v^d)\right|$, i.e. the steps of $u_t$ are lower bounded by a constant, in contradiction to it being bounded. Finally, since $u^t = \sum_{i\in A} w_i$ and those are the only coordinates that diverge, we have that $\mathcal{L}(\mathbf{w}^{(t)}) \leq \exp\left(-\Theta\left(\left\|\mathbf{w}^{(t)}\right\|^k\right)\right)$ □

**Lemma C.3.** *Let $d \geq 2$ be an even integer, and $0 < k < d$ an odd integer. Let $\mathbf{x} = \begin{pmatrix}1\\1\end{pmatrix}, y = 1$, and consider $\mathcal{P}(\mathbf{w}) = \begin{pmatrix}-\left(\sum_{i\in A} w_i\right)^k \\ \left(\sum_{i\in B} w_i\right)^d\end{pmatrix}$, for some $A \cap B = \emptyset$. Further assume that $\left(\sum_{i\in A} w_i\right)^k > 0$ at initialization. Then, optimizing*

$$\mathcal{L}(\mathbf{w}) = \exp\left(-\left(\sum_{i\in A} w_i\right)^k + \left(\sum_{i\in B} w_i\right)^d\right).$$

*With gradient descent yields $\mathcal{L}(\mathbf{w}^{(t)}) = \exp\left(-\Theta\left(\left\|\mathbf{w}^{(t)}\right\|^k\right)\right)$*

*Proof.* The proof is similar to the proof of Lemma C.2, except the boundedness argument, in which case:

- If $k$ is even, $u_t^k \geq 0$ and so $-u_t^k \leq 0 \leq M$
- If $k$ is odd, $u_t^{k-1} \geq 0$, and hence $u_t$ is non-decreasing, so $-u_k^t \leq -u_0^k \leq -M$.

The rest of the proof is similar since we assume $u_0 > 0$. □

# D Effective Degree Characteristics

## D.1 The Effective Degree Depends on the Model, Data, and Initialization

First we formalize Lemma 4.2:

**Lemma D.1.** *Let* $\mathbf{w} = (u, v) \in \mathbb{R}^2$, *define the polynomial predictor* $\mathcal{P}(\mathbf{w}) = \begin{pmatrix} q(\mathbf{w}) \\ q(\mathbf{w})^2 \\ \vdots \\ q(\mathbf{w})^R \end{pmatrix}$, *with*

$$q(\mathbf{w}) = q(u, v) = u^2 + v^4 - u^2 v^2.$$

*For* $r \in [R]$, *let* $\mathbf{x}_r = \mathbf{e}_r$, $y_r = 1$. *Then,* $m_r(\mathbf{w}) = y_r \langle \mathcal{P}(\mathbf{w}), \mathbf{x}_r \rangle = q(\mathbf{w})^r$. *Consider gradient descent with a sufficiently small constant step size on* $\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{R} \exp(-q(\mathbf{w})^r)$. *For the open sets of initializations*

$$U_2 = \{u > 2, |v| < 1/2\}, \qquad U_4 = \{|u| < 1/2, v > 2\}$$
$$\operatorname{effdeg}(m_r) = \begin{cases} 2r & \mathbf{w}^{(0)} \in U_2 \\ 4r & \mathbf{w}^{(0)} \in U_4 \end{cases}.$$

*Proof.* Let us denote $p_2 = u^2$, $p_4 = v^4 - u^2 v^2$ i.e. $q = p_2 + p_4$, these polynomials will be the effective polynomials in the two regimes.

By Lemma D.2 we have that for $\mathbf{w}^{(0)} \in U_2$, $u_t \to \infty$, $v_t \to 0$, $u_t v_t \to 0$, hence $\frac{\mathbf{w}^{(t)}}{\|\mathbf{w}^{(t)}\|} \to \mathbf{e}_1$.

$$\frac{q(\mathbf{w}^{(t)})}{\left\|\mathbf{w}^{(t)}\right\|^2} \to 1 = p_2(\mathbf{e}_1).$$

We have

$$\nabla q(u, v) = (2u(1 - v^2), 2v(2v^2 - u^2)),$$

And so,

$$\frac{\nabla q(\mathbf{w}^{(t)})}{\left\|\mathbf{w}^{(t)}\right\|} \to (2, 0) = \nabla p_2(\mathbf{e}_1).$$

That is, $\operatorname{effdeg}(q) = 2$. In a similar manner, by Lemma D.2 if $\mathbf{w}^{(0)} \in U_4$, then $u_t \to 0$, $v_t = \infty$. So, $\frac{\mathbf{w}^{(t)}}{\|\mathbf{w}^{(t)}\|} \to \mathbf{e}_2$. Furthermore,

$$\frac{q(\mathbf{w}^{(t)})}{\left\|\mathbf{w}^{(t)}\right\|^4} \to 1 = p_4(\mathbf{e}_2),$$

and

$$\frac{\nabla q(\mathbf{w}^{(t)})}{\left\|\mathbf{w}^{(t)}\right\|^3} \to (0, 4) = \nabla p_4(\mathbf{e}_2),$$

that is, $\operatorname{effdeg}(q) = 4$. Finally, recall that $m_r = q^r$. Using the fact that $\frac{m_r(\mathbf{w})}{\|\mathbf{w}\|^{rk}} = \left(\frac{q(\mathbf{w})}{\|\mathbf{w}\|^k}\right)^r$, and $\frac{m_r(\mathbf{w})}{\|\mathbf{w}\|^{rk-1}} = r\left(\frac{q(\mathbf{w})}{\|\mathbf{w}\|^k}\right)^{r-1} \frac{\nabla q(\mathbf{w})}{\|\mathbf{w}\|^{k-1}}$ we finish the proof. □

**Lemma D.2.** *Let* $\mathbf{w} = (u, v) \in \mathbb{R}^2$, *define the polynomial predictor* $\mathcal{P}(\mathbf{w}) = \begin{pmatrix} q(\mathbf{w}) \\ q(\mathbf{w})^2 \\ \vdots \\ q(\mathbf{w})^R \end{pmatrix}$, *with*

$$q(\mathbf{w}) = q(u, v) = u^2 + v^4 - u^2 v^2.$$

*For* $r \in [R]$, *let* $\mathbf{x}_r = \mathbf{e}_r$, $y_r = 1$. *Then,* $m_r(\mathbf{w}) = y_r \langle \mathcal{P}(\mathbf{w}), \mathbf{x}_r \rangle = q(\mathbf{w})^r$. *Consider gradient descent with a sufficiently small constant step size on* $\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{R} \exp(-q(\mathbf{w})^r)$. *For the open sets of initializations*

$$U_2 = \{u > 2, |v| < 1/2\}, \qquad U_4 = \{|u| < 1/2, v > 2\},$$

*and a sufficiently small learning rate* $\eta > 0$, *if* $\mathbf{w}^{(0)} \in U_2$, $\mathbf{w}^{(t)} \in U_2$, $u_t \to \infty$, $v_t \to 0$, $u_t v_t \to 0$. *If* $\mathbf{w}^{(0)} \in U_4$, $\mathbf{w}^{(t)} \in U_4$, $u_t \to 0$, $v_t \to \infty$.

The proof follows by showing the path is invariant to $U_2$ and $U_4$, and the dynamic in each one forces the corresponding asymptotic behavior of the coordinates.

*Proof.* Note that $\nabla m_r(\mathbf{w}) = rq(bw)^{r-1}\nabla q(\mathbf{w})$, hence $\nabla \mathcal{L}(\mathbf{w})$ is always a scalar multiple of $\nabla q(\mathbf{w})$, i.e.

$$-\nabla \mathcal{L}(\mathbf{w}) = a(q(\mathbf{w}))\nabla q(\mathbf{w})$$

For $a(s) = \sum_{r=1}^{R} rs^{r-1}\exp(-s^r)$, which is positive whenever $t > 0$, and in that case we can think of $\mathbf{w}^{(t)}$ following a gradient-ascent path w.r.t. $q$, for appropriate step sizes. The coordinate updates are then:

$$u_{t+1} = u_t\left(1 + 2\eta \cdot a(q(\mathbf{w}^{(t)}))(1 - v_t^2)\right), \qquad v_{t+1} = v_t\left(1 + 2\eta \cdot a(q(\mathbf{w}^{(t)}))(2v_t^2 - u_t^2)\right) \tag{15}$$

Note that $a(s)$ decays exponentially, so we can find

$$M_1 = \sup_{s\geq 3} s \cdot a(s) < \infty, \qquad M_{1/2} = \sup_{s\geq 3} s^{1/2} \cdot a(s) < \infty,$$

We will choose $\eta > 0$ s.t.

$$\frac{8}{3}M_1 < 1/2, \qquad \frac{8}{\sqrt{15}}\eta M_{1/2} < 1/2.$$

**If $\mathbf{w}^{(t)} \in U_2$:** Then $q(\mathbf{w}^{(t)}) = u_t^2(1 - v_t^2) + v_t^4 \geq \frac{3}{4}u_t^2 > 3$, and so $u_t^2 \leq \frac{4}{3}q(\mathbf{w}^{(t)})$. Furthermore, $1 - v_t^2 > 0$, so $u_{t+1} > u_t$, and since $2v_t^2 - u_t^2 < 0$, we have

$$0 > 2\eta \cdot a(q(\mathbf{w}^{(t)}))(2v_t^2 - u_t^2) > -2\eta \cdot a(q(\mathbf{w}^{(t)}))u_t^2 \geq -\frac{8}{3}\eta \cdot a(q(\mathbf{w}^{(t)})) > -1/2,$$

from the choice of $\eta$ and $u_t^2 \leq \frac{4}{3}q(\mathbf{w}^{(t)})$. Plugging in to the update for $v_{t+1}$ we get $|v_{t+1}| < |v_t|$, so $\mathbf{w}^{(t+1)} \in U_2$. We got that $u_t$ is increasing. If we assume it is bounded, we get that $q(\mathbf{w}^{(t)})$ must be bounded, but then we can lower bound $a(q(\mathbf{w}^{(t)})) \geq C > 0$. Together with $1 - v_t^2 \geq 3/4$ we get $u_{t+1} \geq u_t\left(1 + \frac{3}{2}\eta C\right)$, contradiction. So $u_t \to \infty$.

Now we claim that

$$\sum_{i=0}^{\infty} 2\eta a(q(\mathbf{w}^{(t)}))(u_t^2 - 2v_t^2) = \infty \tag{16}$$

. Indeed,

$$\log\frac{u_{t+1}}{u_t} = \log(1 + 2a(q(\mathbf{w}^{(t)}))) \leq 2a(q(\mathbf{w}^{(t)})), \tag{17}$$

and since the left hand side sums to infinity, so must the right hand side. On $U_2$ we have that $u_t^2 - 2v_t^2 > 2/7$, and hence the original sum must diverge.

Now for $\mathbf{w}^{(t)} \in U_2$ we have that

$$\begin{aligned}
0 &< 2\eta a(q(\mathbf{w}^{(t)}))(u_t^2 - 2v_t^2) \\
&< 2\eta a(q(\mathbf{w}^{(t)}))u_t^2 \\
&\leq 2\eta a(q(\mathbf{w}^{(t)}))\frac{4}{3}q(\mathbf{w}^{(t)}) \\
&= \frac{8}{3}q(\mathbf{w}^{(t)})\eta a(q(\mathbf{w}^{(t)})) \\
&\leq \frac{8}{3}\eta M_1 < 1/2.
\end{aligned}$$

Recall that $|v_{t+1}| = |v_t|(1 - 2\eta a(q(\mathbf{w}^{(t)}))(u_t^2 - 2v_t^2))$. Together with Eq. (16) and the bound on the constant multiplying $|v_t|$, we get that $v_t \to 0$. Finally, by Eq. (15)

$$u_{t+1}|v_{t+1}| = u_t|v_t|\left(1 + 2\eta \cdot a(q(\mathbf{w}^{(t)}))(2v_t^2 - u_t^2)\right)\left(1 + 2\eta \cdot a(q(\mathbf{w}^{(t)}))(1 - v_t^2)\right).$$

Using $\log(1+x) \le x$,

$$\log \frac{u_{t+1}|v_{t+1}|}{u_t|v_t|} \le \left(1 + 2\eta \cdot a(q(\mathbf{w}^{(t)}))(2v_t^2 - u_t^2)\right) + 2\eta \cdot a(q(\mathbf{w}^{(t)}))(1 - v_t^2) < -\frac{11}{2}\eta a(q(\mathbf{w}^{(t)}))$$

Where the last inequality is due to $\mathbf{w}^{(t)} \in U_2$. Now summing the above equation over $t$ the r.h.s. diverges to $-\infty$, because the sum over $t$ in Equation (17) diverges to infinity. The bound on the l.h.s. sum yields $|v_t|u_t \to 0$.

**If $\mathbf{w}^{(t)} \in U_4$:** then $u_t^2 < 1/4$, $v_t > 2$. Furthermore,

$$\begin{aligned} 0 &< 2\eta a(q(\mathbf{w}^{(t)}))(v_t^2 - 1) \\ &< 2\eta a(q(\mathbf{w}^{(t)}))v_t^2 \\ &\overset{(*)}{\le} \frac{8}{\sqrt{15}}\eta q(\mathbf{w}^{(t)})^{1/2} a(q(\mathbf{w}^{(t)})) < 1/2. \end{aligned}$$

Where $(*)$ is due to the fact that under $U_4$,

$$q(\mathbf{w}^{(t)}) \ge \frac{15}{16}v_t^4. \tag{18}$$

Recalling the update rule in Equation (15), we get $|u_{t+1}| < |u_t|$, and since $2v_t^2 - u_t^2 > 0$, then $v_{t+1} > v_t$. That is, $\mathbf{w}^{(t+1)} \in U_4$.

Similarly to the prior case, if $v_t$ were bounded then so were $q(\mathbf{w}^{(t)})$, forcing geometric growth for $v_t$, which is a contradiction. Hence $v_t \to \infty$. In a similar technique to the prior case, we would like to show $\sum_{t=0}^{\infty} 2\eta a(q(\mathbf{w}^{(t)}))(v_t^2 - 1)$. First note that $2\eta a(q(\mathbf{w}^{(t)}))(v_t^2 - 1) \ge \frac{2}{3}\eta a(q(\mathbf{w}^{(t)}))v_t^2$, so showing that summing the r.h.s. over $t$ diverges suffices. We have

$$\log \frac{v_{t+1}}{v_t} \le \log\left(1 + 2\eta a(q(\mathbf{w}^{(t)}))(2v_t^2 - u_t^2)\right) \le 4\eta a(q(\mathbf{w}^{(t)}))v_t^2.$$

Since summing the l.h.s. diverges to infinity, so does summing the r.h.s., and hence every scalar multiplication of it. Now $|u_{t+1}| = |u_t|(1 - 2\eta a(q(\mathbf{w}^{(t)}))(v_t^2 - 1))$. From the divergence above and the fact that

$$\begin{aligned} 0 \quad &< \quad 2\eta a(q(\mathbf{w}^{(t)}))(v_t^2 - 1) \\ &\le \quad 2\eta a(q(\mathbf{w}^{(t)}))v_t^2 \\ &\overset{Eq.\ (18)}{\le} \quad \frac{8}{\sqrt{15}}\eta q(\mathbf{w}^{(t)})^{1/2} a(q(\mathbf{w}^{(t)})) \\ &\le \quad \frac{8}{\sqrt{15}}\eta M_{1/2} < 1/2, \end{aligned}$$

we have that $u_t \to 0$. □

### D.2 Determining the Effective Degree Without Solving the Dynamics

We next provide a polynomial predictor $\mathcal{P} : \mathbb{R}^D \to \mathbb{R}^n$, and a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^n$, such that when trained with gradient descent (Equation (1)) with a constant step size $\eta > 0$, the margins $m_i = y_i \langle \mathcal{P}, \mathbf{x}_i \rangle$ have different effective degrees, and effdeg$(m_i)$ can be derived by a short calculation.

Note that the proof works for any subset of the above dataset, with arbitrary labels, and the polynomial predictor can be extended to a family of polynomials for which the above holds. So this yields in fact a family of such examples.

Consider the parameter vector $(u, v) = \mathbf{w} \in \mathbb{R}^2$ and the polynomial $P : \mathbb{R}^2 \to \mathbb{R}^N$

$$P(u,v) = \begin{pmatrix} u^{\nu_1} - v^d \\ u^{\nu_2} - v^d \\ \vdots \\ u^{\nu_N} - v^d \end{pmatrix},$$

for $2 \le \nu_1 < \ldots < \nu_N < d$, with $d$ even. Let $y_i \in \{\pm 1\}$ some labels, choose $\mathbf{x}_i = y_i \mathbf{e}_i \in \mathbb{R}^N$.

Now $m_i(u, v) = u^{\nu_i} - v^d$. For $0 < \eta < \frac{1}{dN}$ and initializing with $u_0 > 1$, and $|v_0| < 1$, we will show $\text{effdeg}(m_i) = \nu_i$. The gradient updates are:

$$\begin{aligned} u_{t+1} &= u_t + \eta \sum_{i=1}^{N} \nu_i u_t^{\nu_i - 1} \exp(v_t^d - u_t^{\nu_i}) \\ v_{t+1} &= v_t - \eta d v_t^{d-1} \sum_{i=1}^{N} \exp(v_t^d - u_t^{\nu_i}) \\ &= v_t \left(1 - \eta d |v_t^{d-2}| \sum_{i=1}^{N} \exp(v_t^d - u_t^{\nu_i})\right) \qquad (d \text{ is even}) \end{aligned}$$

Now while $u_t \ge 1$ and $|v_t| \le 1$, we have $v_t^d - u_t^{\nu_i} \le 0$. So

$$0 \le \eta d |v_t^{d-2}| \sum_{i=1}^{N} \exp(v_t^d - u_t^{\nu_i}) \le \eta d N \le 1$$

So overall $|v_{t+1}| < |v_t|$, and $u_{t+1} \ge u_t$ since all of the $u$ updates are positive.

We have that the optimization stays in the open interval $u_t > u_0 \ge 1$, $|v_t| < |v_0| < 1$.

If we assume $u_t \le M$, then

$$u_{t+1} - u_t \ge \eta \nu_1 u_t^{\nu_1 - 1} \exp\left(v_t^d - u_t^{\nu_1}\right) \ge \eta \nu_1 u_0^{\nu_1 - 1} e^{-M^{\nu_1}} > 0.$$

Hence $u_t \to \infty$, contradiction. A straight forward calculation concludes the proof.

# E Proof of Theorem 5.1

## E.1 Proof Intuition

We have that $\bar{\mathbf{w}}^\infty$ exists from Assumption 2. We use Assumption 1 to show this implies $\bar{\mathbf{w}}^\infty$ must create positive margins, and hence primal feasibility. Assumption 3 implies two different things:

- The margins are dominated by their $\nu_i$-homogeneous parts, and hence $\|\mathcal{P}_i\|$ scale like $\|\mathbf{w}\|^{\nu_i}$.
- The margins' gradients are dominated by their $\nu_i - 1$-homogeneous parts, and hence $\|\nabla \mathcal{P}_i\|$ scale like $\|\mathbf{w}\|^{\nu_i - 1}$.

The first is then used to transform the non-homogeneous margin $m_i := y_i \langle \mathcal{P}_i, \mathbf{x}_i \rangle$ into the homogeneous polynomial in the KKT conditions, i.e. $m^{\nu_i} = \left\langle y_i C_i \mathcal{P}_i^{\nu_i}, \mathbf{x}_i \right\rangle$, and also to show that the gradients of lower-effective-degree margins dominate the gradient, since their corresponding loss-exponent decays slower (the exponent of every loss-per-sample decays as $\|\mathbf{w}\|^{\nu_i}$). Among the lowest effdeg margins, we further show that those that converge to a lower normalized margin dominate their higher-normalized-margin counterparts. Overall this tells us that $\nabla \mathcal{L} = \sum_{i \in S} \alpha_i \nabla m_i + o\left(\sum_{i \in S} \alpha_i \nabla m_i\right)$, where $S$ consists of the lowest effective degree polynomials with the lowest normalized margin, i.e. the gradient is dominated by a possibly dynamic convex sum of margins from $S$. From Assumption 3, $\nabla m_i \approx C_i \nabla m_i^{\nu_Z}$ where $\nu_Z$ is the minimal effective degree and hence $(\bar{\mathbf{w}}^\infty)^\top \sum_{i \in S} \alpha_i \nabla m_i \approx \gamma_{\min} \nu_Z$, where $\gamma_{\min}$ is the minimum margin in that defines $S$. i.e. the loss has a constant projection onto $\bar{\mathbf{w}}^\infty$. We then attempt to show that the projection onto the space orthogonal to $\bar{\mathbf{w}}^\infty$ vanishes, at least in some subsequence of the gradients. Towards this we show that the GD updates induce an approximate exponential-multiplicative-weights step on $\alpha_i$, which can be thought of as the dual variables. We further show that this multiplicative-weights update is one that minimizes the gradient norm along some subsequence. Finally, we show that the minimizer must have no component orthogonal to $\bar{\mathbf{w}}^\infty$. This implies that there exists a subsequence for which $\nabla \mathcal{L} \propto \bar{\mathbf{w}}^\infty$, which shows optimality and complementary slackness. Dual feasibility is due to $\alpha_i$ being weights of a convex sum.

## E.2 Proof

For convenience, let us restate the theorem:

**Theorem E.1.** *Let $\{\mathcal{P}_i\}_{i=1}^N$ a collection of polynomials with parameters $\mathbf{w} \in \mathbb{R}^D$. For a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^N$ where $y_i = \operatorname{sgn}(\langle \mathcal{P}_i(\mathbf{w}), \mathbf{x}_i \rangle)$, consider the sequence of gradient updates (Equation (1)) with a non-increasing sequence of step sizes $\{\eta_t\}_t$, minimizing the empirical risk. If Assumptions 1 to 3 hold we have that $\bar{\mathbf{w}}^\infty := \lim_{t\to\infty} \frac{\mathbf{w}^{(t)}}{\|\mathbf{w}^{(t)}\|}$ is a positive scaling of a first order stationary point (KKT point) of the following optimization problem:*

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w}\|_2^2 \ \text{s.t.}\ \forall i, y_i \left\langle \mathbf{x}_i, \mathcal{P}_i^{\nu_i}(\mathbf{w}) \right\rangle \geq 0 \\ \forall i \in Z, y_i C_i \left\langle \mathbf{x}_i, \mathcal{P}_i^{\nu_i}(\mathbf{w}) \right\rangle \geq 1 \,, \end{aligned} \tag{OPT}$$

*where $Z = \arg\min_{i\in[n]} \left\{ \operatorname{effdeg}_{\mathbf{w}} y_i \langle \mathcal{P}_i, \mathbf{x}_i \rangle \right\}$.*

We would like to show that $\bar{\mathbf{w}}^\infty$ is in a direction of a KKT point of Equation (OPT). First, let us state the KKT conditions:

- Primal feasibility: $\forall i, y_i \left\langle \mathbf{x}_i, \mathcal{P}_i^{\nu_i}(\mathbf{w}) \right\rangle \geq 0, \forall i \in Z, y_i C_i \left\langle \mathbf{x}_i, \mathcal{P}_i^{\nu_i}(\mathbf{w}) \right\rangle \geq 1$
- Dual feasibility and complementary slackness: There are $\{\mu_i\}_{i=1}^N$ where $\mu_i \geq 0$ and
  - $\forall i \notin Z$, if $y_i \left\langle \mathbf{x}_i, \mathcal{P}_i^{\nu_i}(\mathbf{w}) \right\rangle > 0$, $\mu_i = 0$
  - $\forall i \in Z$, if $y_i C_i \left\langle \mathbf{x}_i, \mathcal{P}_i^{\nu_i}(\mathbf{w}) \right\rangle > 1$, $\mu_i = 0$
- Optimality: $\mathbf{w} = \sum_{i=1}^N \mu_i \cdot C_i \nabla \mathcal{P}_i^{\nu_i}(\mathbf{w})^\top y_i \mathbf{x}_i$

First we will show that any positive scaling of $\bar{\mathbf{w}}^\infty$ is a feasible solution.

**Lemma E.2** (Primal Feasibility)**.** *$\bar{\mathbf{w}}^\infty$ classifies all examples correctly, and for $C > 0$ large enough, $C\bar{\mathbf{w}}^\infty$ is a feasible solution to Equation (OPT) under Assumptions 1 to 3.*

*Proof.* By Assumption 3 we have that $\frac{\mathcal{P}_i(\mathbf{w}^{(t)})^\top \mathbf{x}_i}{\|\mathbf{w}^{(t)}\|^{\nu_i}} - \delta_i^{(t)} = C_i \mathcal{P}_i^{\nu_i}(\bar{\mathbf{w}}^\infty)^\top \mathbf{x}_i \neq 0$ with $\delta_i^{(t)} \to 0$. Hence from homogeneity $C_i \mathcal{P}_i^{\nu_i}(C\bar{\mathbf{w}}^\infty)^\top \mathbf{x}_i \neq 0$. Assume towards contradiction that $C_i \mathcal{P}_i^{\nu_i}(C\bar{\mathbf{w}}^\infty)^\top \mathbf{x}_i y_i < 0$, then we have

$$\mathcal{P}_i(\mathbf{w}^{(t)})^\top \mathbf{x}_i y_i = \left( \frac{C_i \mathcal{P}_i^{\nu_i}(C\bar{\mathbf{w}}^\infty)^\top \mathbf{x}_i y_i}{C^{\nu_i}} + \delta_i^{(t)} \right) \left\| \mathbf{w}^{(t)} \right\|^{\nu_i} \to -\infty$$

Since $\delta_i^{(t)} \to 0$, contradicting Assumption 1. All that is left is scaling $C$ such that all margins are in the feasible region. □

To show the rest of the KKT conditions, we want to understand the asymptotic direction of the gradient.

**Lemma E.3** (Minimum Margin of Minimum Effective Degree Dominates the Gradient)**.** $-\overline{\nabla \mathcal{L}(\bar{\mathbf{w}}^{(t)})} + \delta_\nabla^{(t)} \propto \sum_{i\in S} \alpha_i^{(t)} C_i \nabla \mathcal{P}_i^{\nu_i} \left(\mathbf{w}^{(t)}\right)^\top \mathbf{x}_i y_i$ *for $\delta_\nabla^{(t)} \to 0$, where $S = \arg\min_{i\in Z} \{C_i \mathcal{P}_i^{\nu_Z}(\bar{\mathbf{w}}^\infty)^\top \mathbf{x}_i y_i\}$, where $S$ is a set, and $Z = \arg\min_{i\in[n]} \left\{ \operatorname{effdeg}\left(\mathcal{P}_i^\top(\bar{\mathbf{w}}^\infty)\mathbf{x}_i y_i\right) \right\}$.*

*Proof.* Consider the gradient of the loss:

$$\begin{aligned} -\nabla \mathcal{L}\left(\mathbf{w}^{(t)}\right) &= \sum_{i=1}^n \nabla \mathcal{P}_i \left(\mathbf{w}^{(t)}\right)^\top \mathbf{x}_i y_i \exp\left(-\mathcal{P}_i \left(\mathbf{w}^{(t)}\right)^\top \mathbf{x}_i y_i\right) \\ &= \sum_{i=1}^n \nabla \mathcal{P}_i \left(\mathbf{w}^{(t)}\right)^\top \mathbf{x}_i y_i \exp\left(-g^{\nu_i}(t) \left(C_i \mathcal{P}_i^{\nu_i} \left(\bar{\mathbf{w}}^\infty\right)^\top \mathbf{x}_i y_i + \delta_i^{(t)} y_i\right)\right) \quad \text{(by assumption 3)} \end{aligned}$$

Now denoting $\gamma_i := C_i \mathcal{P}_i^{\nu_i} \left(\bar{\mathbf{w}}^\infty\right)^\top \mathbf{x}_i y_i > 0$ (this is positive from Lemma E.2), we get

$$-\nabla \mathcal{L}\left(\mathbf{w}^{(t)}\right) = \sum_{i=1}^n \nabla \mathcal{P}_i \left(\mathbf{w}^{(t)}\right)^\top \mathbf{x}_i y_i \exp\left(-g^{\nu_i}(t) \left(\gamma_i + \delta_i^{(t)} y_i\right)\right)$$

Now we want to divide the loss into two separate parts. Denote $\nu_Z := \min_i\{\text{effdeg}\,(\mathcal{P}_i(\cdot)^\top \mathbf{x}_i)\}$ ($Z$ is chosen because for every $k \in Z$ effdeg $(\mathcal{P}_i(\cdot)^\top \mathbf{x}_i) = \nu_Z$). Let us write

$$A(t) = \sum_{i\in Z} \nabla\mathcal{P}_i\left(\mathbf{w}^{(t)}\right)\mathbf{x}_i y_i^\top \exp\left(-g^{\nu_Z}(t)\left(\gamma_i + \delta_i^{(t)}\right)\right)$$

and

$$B(t) = \sum_{i\notin Z} \nabla\mathcal{P}_i\left(\mathbf{w}^{(t)}\right)\mathbf{x}_i y_i^\top \exp\left(-g^{\nu_i}(t)\left(\gamma_i + \delta_i^{(t)}\right)\right)$$

Note that $A(t) + B(t) = -\nabla\mathcal{L}(\mathbf{w}^{(t)})$. We will now further divide $A(t)$ into two parts. First denote $\gamma_{\min} := \min_i \gamma_i$, (recall they are all equal and minimal), and let $\frac{\min_{i\in Z\setminus S}\{\gamma_i\}+\gamma_{\min}}{2} = \hat{\gamma}$. The idea is that we define $\hat{\gamma}$ to be the average between the minimum margin (of the minimum effective degree), and the second-smallest margin (of the minimum effdeg). Then, when we will have for $i \in Z \setminus S$ $\gamma_i - \hat{\gamma} > 0$ while $\gamma_{\min} - \hat{\gamma} < 0$. Concretely, denote:

$$\mathrm{I}(t) := \frac{\sum_{i\in S}\exp\left(-g^{\nu_Z}(t)\left(\gamma_{\min} + \delta_i^{(t)}\right)\right)\nabla\mathcal{P}_i(\mathbf{w}^{(t)})^\top\mathbf{x}_i y_i}{\exp\left(-\hat{\gamma}\cdot g^{\nu_Z}(t)\right)}$$

$$\mathrm{II}(t) := \frac{\sum_{i\in Z\setminus S}\exp\left(-g^{\nu_Z}(t)\left(\gamma_i + \delta_i^{(t)}\right)\right)\nabla\mathcal{P}_i(\mathbf{w}^{(t)})^\top\mathbf{x}_i y_i}{\exp\left(-\hat{\gamma}\cdot g^{\nu_Z}(t)\right)}$$

We chose I, II such that $A(t) = \exp(-\hat{\gamma}\cdot g^{\nu_Z}(t))(\mathrm{I}(t) + \mathrm{II}(t))$. We can then write

$$\|A(t)\| \geq \exp(-\hat{\gamma}\cdot g^{\nu_Z}(t))\cdot\left|\|\mathrm{I}(t)\| - \|\mathrm{II}(t)\|\right| \tag{19}$$

Inspecting the norm

$$\|\mathrm{I}(t)\| = \left\|\sum_{i\in S}\exp\left(-g^{\nu_Z}(t)\left(\gamma_{\min} - \hat{\gamma} + \delta_i^{(t)}\right)\right)\nabla\mathcal{P}_i(\mathbf{w}^{(t)})^\top\mathbf{x}_i y_i\right\|$$

We note that for large enough $t$ since $\delta_i^{(t)} \to 0$ and by definition of $\hat{\gamma}$

$$\gamma_{\min} - \hat{\gamma} + \delta_i^{(t)} < 0 \tag{20}$$

We want to use this to show $\|\mathrm{I}(t)\| \to \infty$. Here, a subtle point is to verify the linear combination doesn't asymptotically cancel out. Indeed

$$\begin{aligned}\mathrm{I}(t) &= \sum_{i\in S}\exp\left(-g^{\nu_Z}(t)\left(\gamma_{\min} - \hat{\gamma} + \delta_i^{(t)}\right)\right)\nabla\mathcal{P}_i(\mathbf{w}^{(t)})^\top\mathbf{x}_i y_i \\ &= \sum_{i\in S}\exp\left(-g^{\nu_Z}(t)\left(\gamma_{\min} - \hat{\gamma} + \delta_i^{(t)}\right)\right)\left(C_i\nabla\mathcal{P}_i^{\nu_Z}(\bar{\mathbf{w}}^\infty)^\top\mathbf{x}_i y_i + \delta_{\nabla,i}^{(t)}\right) \qquad \text{(Assumption 3)}\end{aligned}$$

Now from homogeneity we have that

$$(\bar{\mathbf{w}}^\infty)^\top\, C_i\nabla\mathcal{P}_i^{\nu_Z}(\bar{\mathbf{w}}^\infty)y_i\mathbf{x}_i = \nu_Z C_i\mathcal{P}_i^{\nu_Z}(\bar{\mathbf{w}}^\infty)y_i\mathbf{x}_i = \nu_Z\gamma_i > 0$$

Hence

$$\frac{(\bar{\mathbf{w}}^\infty)^\top\mathrm{I}(t)}{\|\mathrm{I}(t)\|} = \frac{\sum_{i\in S}\exp\left(-g^{\nu_Z}(t)\left(\gamma_{\min} - \hat{\gamma} + \delta_i^{(t)}\right)\right)}{\left\|\sum_{i\in S}\exp\left(-g^{\nu_Z}(t)\left(\gamma_{\min} - \hat{\gamma} + \delta_i^{(t)}\right)\right)\left(C_i\nabla\mathcal{P}_i^{\nu_Z}(\bar{\mathbf{w}}^\infty)^\top\mathbf{x}_i y_i + \delta_{\nabla,i}^{(t)}\right)\right\|}\left(\nu_Z\gamma_{\min} + \delta_{\nabla,i}^{(t)}\right)$$

Is bounded away from 0, and so is $\frac{\mathrm{I}(t)}{\|\mathrm{I}(t)\|}$. Together with Equation (20), this yields $\|\mathrm{I}(t)\| \to \infty$.

Now for $\mathrm{II}(t)$,

$$\|\mathrm{II}(t)\| = \left\|\frac{\sum_{i\in Z\setminus S}\exp\left(-g^{\nu_Z}(t)\left(\gamma_i + \delta_i^{(t)}\right)\right)\nabla\mathcal{P}_i(\mathbf{w}^{(t)})^\top\mathbf{x}_i y_i}{\exp\left(-\hat{\gamma}\cdot g^{\nu_Z}(t)\right)}\right\|$$

here again by the definition of $\hat{\gamma}$ and for large enough $t$, we have that for any $i \in Z \setminus S$

$$\gamma_i - \hat{\gamma} + \delta_i^{(t)} > 0$$

therefore we can bound $\|\mathrm{II}(t)\| \leq C$ for large enough $t$. Now consider the expression

$$\|B(t)\| = \left\| \sum_{i \notin Z} \nabla \mathcal{P}_i \left( \mathbf{w}^{(t)} \right) \mathbf{x}_i y_i^\top \exp\left( -g^{\nu_i}(t) \left( \gamma_i + \delta_i^{(t)} \right) \right) \right\| \tag{21}$$

$$\leq \sum_{i \notin Z} \exp\left( -g^{\nu_i}(t) \left( \gamma_i + \delta_i^{(t)} \right) \right) \left\| \nabla \mathcal{P}_i \left( \mathbf{w}^{(t)} \right) \mathbf{x}_i \right\| \tag{22}$$

$$\leq \sum_{i \notin Z} g^{\nu_i - 1}(t) \exp\left( -g^{\nu_i}(t) \left( \gamma_i + \delta_i^{(t)} \right) \right) (r_{\max} + C) \tag{23}$$

For some constant $C > 0$, and $r_{\max} = \max_i \left\| C_i \nabla \mathcal{P}_i^{\nu_i} (\bar{\mathbf{w}}^\infty)^\top \mathbf{x}_i \right\|$. Overall using the bound on $\|A(t)\|$ (19), we get that

$$\frac{\|B(t)\|}{\|A(t)\|} \leq \frac{\sum_{i \notin Z} g^{\nu_i - 1}(t) \exp\left( -g^{\nu_i}(t) \left( \gamma_i + \delta_i^{(t)} \right) \right) (r_{\max} + C)}{\exp(-\hat{\gamma} g^{\nu_Z}(t)) \cdot |\|\mathrm{I}(t)\| - \|\mathrm{II}(t)\||}$$

$$= \frac{\sum_{i \notin Z} g^{\nu_i - 1}(t) \exp\left( -g^{\nu_i}(t) \left( \gamma_i + \delta_i^{(t)} \right) + \hat{\gamma} g^{\nu_Z}(t) \right) (r_{\max} + C)}{|\|\mathrm{I}(t)\| - \|\mathrm{II}(t)\||}$$

Now for $i \notin Z$, $-g^{\nu_i}(t) \left( \gamma_i + \delta_i^{(t)} \right) + \hat{\gamma} g^{\nu_Z}(t)$ is negative for large enough $t$. Since $\hat{\gamma}, \gamma_i + \delta_i^{(t)}$ are bounded and positive for large enough $t$, and $g^{\nu_i}(t)$ grows asymptotically faster than $g^{\nu_Z}(t)$. Therefore the expression in the exponent is negative for large enough $t$, hence the numerator goes to zero. The denominator goes to infinity due to $\|\mathrm{I}(t)\| \to \infty$, and that $\|\mathrm{II}(t)\|$ is bounded, so overall:

$$-\frac{\nabla \mathcal{L}(\mathbf{w}^{(t)})}{\left\| \nabla \mathcal{L}(\mathbf{w}^{(t)}) \right\|} = \frac{A(t) + B(t)}{\|A(t) + B(t)\|} = \frac{A(t)}{\|A(t)\|} + \tilde{\delta}^{(t)}$$

For some $\tilde{\delta}^{(t)} \to 0$. Now similarly consider $\frac{\exp(-\hat{\gamma} g^{\nu_Z}(t)) \cdot \|\mathrm{II}(t)\|}{\exp(-\hat{\gamma} g^{\nu_Z}(t)) \cdot \|\mathrm{I}(t)\|} = \frac{\|\mathrm{II}(t)\|}{\|\mathrm{I}(t)\|} \to 0$, hence

$$\frac{A(t)}{\|A(t)\|} = \frac{\mathrm{I}(t)}{\|\mathrm{I}(t)\|} + \hat{\delta}^{(t)}$$

For $\hat{\delta}^{(t)} \to 0$. Overall

$$-\frac{\nabla \mathcal{L}(\mathbf{w}^{(t)})}{\left\| \nabla \mathcal{L}(\mathbf{w}^{(t)}) \right\|} \to \frac{\mathrm{I}(t)}{\|\mathrm{I}(t)\|}$$

i.e. the normalized gradient is dominated by $\mathrm{I}(t)$. □

For simplicity of notation, let us denote $m_i(\mathbf{w}) := \mathcal{P}_i(\mathbf{w})^\top \mathbf{x}_i y_i$. Let us define $\forall i \in S, \alpha_i^{(t)} := \frac{\exp(-m_i(\mathbf{w}^{(t)}))}{\mathcal{L}_{|S}(\mathbf{w}^{(t)})}$, where $\mathcal{L}_{|S} = \sum_{i \in S} \exp(-m_i) \leq \mathcal{L}$. By Assumption 3, we have that $\nabla m_i(\mathbf{w}^{(t)}) = \left\| \mathbf{w}^{(t)} \right\|^{\nu_i} \left( C_i \nabla m_i^{\nu_i}(\bar{\mathbf{w}}^\infty) + o(1) \right)$. So overall

$$-\nabla \mathcal{L}(\mathbf{w}^{(t)}) = \mathcal{L}_{|S}(\mathbf{w}^{(t)}) g^{\nu_Z - 1}(t) \left( \sum_{i \in S} \alpha_i^{(t)} C_i \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^\infty) + o(1) \right)$$

Now note that $\Delta \mathbf{w}^{(t)} := \mathbf{w}^{(t+1)} - \mathbf{w}^{(t)} = -\eta_t \nabla \mathcal{L}(\mathbf{w}^{(t)}) \to 0$, hence using taylor approximation we have

$$m_i(\mathbf{w}^{(t+1)}) = m_i(\mathbf{w}^{(t)}) + \left\langle \nabla m_i(\mathbf{w}^{(t)}), \Delta \mathbf{w}^{(t)} \right\rangle + E_i^{(t)}$$

With

$$\begin{aligned}\left|E_i^{(t)}\right| &\leq K_i \left\|\Delta \mathbf{w}^{(t)}\right\|^2 g^{\deg(m_i)-2}(t) \\ &= K_i \eta_t^2 \left\|\nabla \mathcal{L}(\mathbf{w}^{(t)})\right\|^2 g^{\deg(m_i)-2}(t) \\ &\leq K_i B^2 \eta_t^2 \mathcal{L}_{|S}^2(\mathbf{w}^{(t)}) g^{2\nu_Z+\deg(m_i)-4}(t)\end{aligned}$$

For some constants $K_i$, and for some $B$ that bounds $\left(\sum_{i\in S} \alpha_i^{(t)} C_i \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^\infty) + o(1)\right)$. Now let us compute the inner product from the above expression:

$$\left\langle \nabla m_i(\mathbf{w}^{(t)}), \Delta \mathbf{w}^{(t)} \right\rangle = \eta_t \sum_j \exp(-m_j(\mathbf{w}^{(t)})) \left\langle \nabla m_i(\mathbf{w}^{(t)}), \nabla m_j(\mathbf{w}^{(t)}) \right\rangle$$

So we have for $i \in S$

$$m_i(\mathbf{w}^{(t+1)}) - m_i(\mathbf{w}^{(t)}) = \eta_t \mathcal{L}_{|S}(\mathbf{w}^{(t)}) g^{2\nu_Z-2}(t) \left( \sum_j \alpha_j^{(t)} \left\langle C_i \nabla m_i^{\nu_i}(\bar{\mathbf{w}}^\infty), C_j \nabla m_j^{\nu_j}(\bar{\mathbf{w}}^\infty) \right\rangle \right) + o(\eta_t \mathcal{L}_{|S}(\mathbf{w}^{(t)}) g^{2\nu_Z-2}(t))$$

Denote $\beta^{(t)} := \eta_t \mathcal{L}_{|S}(\mathbf{w}^{(t)}) g^{2\nu_Z-2}(t)$, $A_{i,j} := \left\langle C_i \nabla m_i^{\nu_i}(\bar{\mathbf{w}}^\infty), C_j \nabla m_j^{\nu_j}(\bar{\mathbf{w}}^\infty) \right\rangle$ a gram matrix, then we can re-write $m_i(\mathbf{w}^{(t+1)}) - m_i(\mathbf{w}^{(t)}) = \beta^{(t)}(A\alpha^{(t)})_i + o(\beta^{(t)})$. Exponentiating we get

$$\exp(-m_i(\mathbf{w}^{(t+1)}) + m_i(\mathbf{w}^{(t)})) = \exp(-\beta^{(t)}(A\alpha^{(t)})_i + o(\beta^{(t)})) \tag{24}$$

Now, let us calculate the $\alpha_i^{(t)}$ update:

$$\begin{aligned}\alpha_i^{(t+1)} &= \frac{\exp(-m_i(\mathbf{w}^{(t+1)}))}{\sum_{j\in S} \exp(-m_j(\mathbf{w}^{(t+1)}))} \\ &= \frac{\exp(-m_i(\mathbf{w}^{(t)})) \exp(-m_i(\mathbf{w}^{(t+1)}) + m_i(\mathbf{w}^{(t)}))}{\sum_{j\in S} \exp(-m_j(\mathbf{w}^{(t)})) \exp(-m_j(\mathbf{w}^{(t+1)}) + m_j(\mathbf{w}^{(t)}))} && \text{(adding \& subtracting)} \\ &= \frac{\exp(-m_i(\mathbf{w}^{(t)})) \left(\exp(-\beta^{(t)}(A\alpha^{(t)})_i + o(\beta^{(t)})\right)}{\sum_{j\in S} \exp(-m_j(\mathbf{w}^{(t)})) \left(\exp(-\beta^{(t)}(A\alpha^{(t)})_j + o(\beta^{(t)})\right)} && \text{(by Equation (24))} \\ &= \frac{\alpha_i^{(t)} \left(\exp(-\beta^{(t)}(A\alpha^{(t)})_i + o(\beta^{(t)})\right)}{\sum_{j\in S} \alpha_j^{(t)} \left(\exp(-\beta^{(t)}(A\alpha^{(t)})_j + o(\beta^{(t)})\right)} && (\exp(-m_i(\mathbf{w}^{(t)})) = \alpha_i^{(t)} \mathcal{L}_{|S}(\mathbf{w}^{(t)}))\end{aligned} \tag{25}$$

Note that this is an Normalized Exponentiated Gradient update with perturbations.

**Lemma E.4.** *Let $\alpha^{(t)}$ be a sequence of vectors on the simplex, let $\alpha^*$ be some vector on the simplex, and let $\Phi(\alpha)$ be some convex function to be optimized. Denote $\ell^{(t)} = \nabla\Phi(\alpha^{(t)})$ and assume*

$$\alpha_i^{(t+1)} = \frac{\alpha_i^{(t)} \exp\left(-\beta^{(t)} \ell_i^{(t)} + \varepsilon_i^{(t)}\right)}{\sum_j \alpha_j^{(t)} \exp\left(-\beta^{(t)} \ell_j^{(t)} + \varepsilon_j^{(t)}\right)}$$

*For $\varepsilon^{(t)} \to 0$, then,*

$$\mathrm{KL}\left(\alpha^* \parallel \alpha^{(t+1)}\right) - \mathrm{KL}\left(\alpha^* \parallel \alpha^{(t)}\right) \leq \beta^{(t)} \left(\Phi(\alpha^*) - \Phi(\alpha^{(t)})\right) + (\beta^{(t)}\Delta^{(t)})^2/8 + 2\left\|\varepsilon^{(t)}\right\|_\infty \tag{26}$$

*Where $\Delta^{(t)} := \max_i \ell_i^{(t)} - \min_i \ell_i^{(t)}$.*

*Proof.* First, we will show

$$\mathrm{KL}\left(\alpha^* \parallel \alpha^{(t+1)}\right) - \mathrm{KL}\left(\alpha^* \parallel \alpha^{(t)}\right) \leq \beta^{(t)} \left\langle \ell^{(t)}, \alpha^* - \alpha^{(t)} \right\rangle + (\beta^{(t)}\Delta^{(t)})^2/8 + 2\left\|\varepsilon^{(t)}\right\|_\infty \tag{27}$$

For some generic $\ell^{(t)}$. Denote $Z_t := \sum_j \alpha_j^{(t)} \exp\left(-\beta^{(t)}\ell_i^{(t)} + \varepsilon_i^{(t)}\right)$, so

$$\alpha_i^{(t+1)} = \frac{\alpha_i^{(t)} \exp\left(-\beta^{(t)}\ell_i^{(t)} + \varepsilon_i^{(t)}\right)}{Z_t}$$

Taking log over the update equation, we get $\log \alpha_i^{(t+1)} = \log(\alpha_i^{(t)}) - \beta^{(t)}\ell_i^{(t)} + \varepsilon_i^{(t)} - \log(Z_t)$ Hence,

$$\log \frac{\alpha_i^{(t)}}{\alpha_i^{(t+1)}} = \beta^{(t)}\ell_i^{(t)} - \varepsilon_i^{(t)} + \log(Z_t) \tag{28}$$

So overall,

$$\begin{aligned}
\mathrm{KL}\left(\alpha^* \parallel \alpha^{(t+1)}\right) - \mathrm{KL}\left(\alpha^* \parallel \alpha^{(t)}\right) &= \sum_i \alpha_i^* \log \frac{\alpha_i^*}{\alpha_i^{(t+1)}} - \sum_i \alpha_i^* \log \frac{\alpha_i^*}{\alpha_i^{(t)}} \\
&= \sum_i \alpha_i^* \left(\log \frac{\alpha_i^*}{\alpha_i^{(t+1)}} + \log \frac{\alpha_i^{(t)}}{\alpha_i^*}\right) \\
&= \sum_i \alpha_i^* \log \frac{\alpha_i^{(t)}}{\alpha_i^{(t+1)}} \\
&= \sum_i \alpha_i^* \left(\beta^{(t)}\ell_i^{(t)} - \varepsilon_i^{(t)} + \log(Z_t)\right) \qquad \text{(eq. (28))} \\
&= \beta^{(t)} \left\langle \alpha^*, \ell^{(t)} \right\rangle - \left\langle \alpha^*, \varepsilon^{(t)} \right\rangle + \log Z_t \qquad (29)
\end{aligned}$$

Now let us bound $\log Z_t$. Denote $\mu_t := \left\langle \alpha^{(t)}, \ell^{(t)} \right\rangle$ (We can think as $\mu_t = \mathbb{E}_i \ell_i^{(t)}$, where $i$ is distributed by $\alpha^{(t)}$ – a simplex vector). Then adding and subtracting $\mu_t$ we get:

$$Z_t = \sum_i \alpha_i^{(t)} \exp(-\beta^{(t)}(\ell_i^{(t)} - \mu_t)) \exp(\varepsilon_i^{(t)}) \exp(-\beta^{(t)}\mu_t)$$

Hence,

$$\begin{aligned}
\log Z_t &= -\beta^{(t)}\mu_t + \log \sum_i \alpha_i^{(t)} \exp(-\beta^{(t)}(\ell_i^{(t)} - \mu_t)) \exp(\varepsilon_i^{(t)}) \\
&\le -\beta^{(t)}\mu_t + \log\left(\max_i \exp(\varepsilon_i^{(t)}) \sum_i \alpha_i^{(t)} \exp(-\beta^{(t)}(\ell_i^{(t)} - \mu_t))\right) \\
&= -\beta^{(t)}\mu_t + \left\|\varepsilon^{(t)}\right\|_\infty + \log\left(\sum_i \alpha_i^{(t)} \exp(-\beta^{(t)}(\ell_i^{(t)} - \mu_t))\right)
\end{aligned}$$

Now again, since $\alpha^{(t)}$ is a simplex vector (or a probability distribution) we can use Hoeffding's Lemma (Lemma F.3), and get $\log\left(\sum_i \alpha_i^{(t)} \exp(-\beta^{(t)}(\ell_i^{(t)} - \mu_t))\right) \le \frac{\left(\beta^{(t)}\Delta^{(t)}\right)^2}{8}$. Hence,

$$\log Z_t \le -\beta^{(t)}\mu_t + \left\|\varepsilon^{(t)}\right\|_\infty + \frac{\left(\beta^{(t)}\Delta^{(t)}\right)^2}{8}$$

Plugging it in Equation (29) we get

$$\begin{aligned}
\mathrm{KL}\left(\alpha^* \parallel \alpha^{(t+1)}\right) - \mathrm{KL}\left(\alpha^* \parallel \alpha^{(t)}\right) &\le \beta^{(t)} \left\langle \ell^{(t)}, \alpha^* - \alpha^{(t)} \right\rangle + \left\|\varepsilon^{(t)}\right\|_\infty + \frac{\left(\beta^{(t)}\Delta^{(t)}\right)^2}{8} - \left\langle \alpha^*, \varepsilon^{(t)} \right\rangle \\
&\le \beta^{(t)} \left\langle \ell^{(t)}, \alpha^* - \alpha^{(t)} \right\rangle + 2\left\|\varepsilon^{(t)}\right\|_\infty + \frac{\left(\beta^{(t)}\Delta^{(t)}\right)^2}{8}
\end{aligned}$$

Finally, take $\ell^{(t)} = \nabla\Phi(\alpha^{(t)})$. From the convexity of $\Phi$ we get $\Phi(\alpha^*) - \Phi(\alpha^{(t)}) \ge \left\langle \nabla\Phi(\alpha^{(t)}), \alpha^* - \alpha^{(t)} \right\rangle = \left\langle \ell^{(t)}, \alpha^* - \alpha^{(t)} \right\rangle$. Substitute to get what's required. □

**Lemma E.5.** *1.* $\beta^{(t)} \to 0$

*2.* $\sum_{t=0}^{\infty} \beta^{(t)} = \infty$

*Proof.* For Item 1, recall that $\beta^{(t)} = \eta_t \mathcal{L}_{|S}(\mathbf{w}^{(t)}) g^{2\nu_Z - 2}(t)$. So while $g^{2\nu_Z-2}(t)$ is polynomial in $\left\|\mathbf{w}^{(t)}\right\| = g(t)$, $\mathcal{L}_{|S}(\mathbf{w}^{(t)})$ is exponentially decaying in $g(t)$. Since $g(t) \to \infty$ we have $\beta^{(t)} \to 0$. For Item 2 let us analyze the radial increments $\Delta g^{\nu_Z}(t) := g^{\nu_Z}(t+1) - g^{\nu_Z}(t)$. Recall that $\Delta \mathbf{w}^{(t)} = -\eta_t \nabla \mathcal{L}(\mathbf{w}^{(t)})$. Using tailor expansion on the norm we get

$$g(t+1) = \left\|\mathbf{w}^{(t)} + \Delta \mathbf{w}^{(t)}\right\| = g(t) + \left\langle \bar{\mathbf{w}}^{(t)}, \Delta \mathbf{w}^{(t)} \right\rangle + O\left(\frac{\left\|\Delta \mathbf{w}^{(t)}\right\|^2}{g(t)}\right) \tag{30}$$

Now using taylor expansion on $g(t) \mapsto g(t)^{\nu_Z}$ for a small increment

$$\begin{aligned} g(t+1)^{\nu_Z} &= g(t)^{\nu_Z} + \nu_Z g(t)^{\nu_Z - 1}(g(t+1) - g(t)) + O\left(g(t)^{\nu_Z-2}(g(t+1)-g(t))^2\right) \\ &= g(t)^{\nu_Z} + \nu_Z g(t)^{\nu_Z-1} \left\langle \bar{\mathbf{w}}^{(t)}, \Delta \mathbf{w}^{(t)} \right\rangle + O\left(g(t)^{\nu_Z-2}\left\|\Delta \mathbf{w}^{(t)}\right\|^2\right) && \text{(by Equation (30))} \end{aligned}$$

Now note that

$$\begin{aligned} \nu_Z g(t)^{\nu_Z-1} \left\langle \bar{\mathbf{w}}^{(t)}, \Delta \mathbf{w}^{(t)} \right\rangle &= \eta_t \mathcal{L}_{|S}(\mathbf{w}^{(t)}) \nu_Z g(t)^{2\nu_Z-2} \left( \left\langle \frac{\mathbf{w}^{(t)}}{\left\|\mathbf{w}^{(t)}\right\|}, \sum_{i\in S} \alpha_i^{(t)} C_i \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^{\infty}) \right\rangle + o(1) \right) \\ &= \beta^{(t)} \nu_Z \left( \left\langle \frac{\mathbf{w}^{(t)}}{\left\|\mathbf{w}^{(t)}\right\|}, \sum_{i\in S} \alpha_i^{(t)} C_i \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^{\infty}) \right\rangle + o(1) \right) && (\text{def } \beta^{(t)}) \\ &\le \beta^{(t)} \nu_Z \left( \left\| \sum_{i\in S} \alpha_i^{(t)} C_i \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^{\infty}) \right\| + o(1) \right) && \text{(Cauchy-Schwartz)} \\ &\le \beta^{(t)} \nu_Z C_M \end{aligned}$$

For $C_M = 1 + \max_{i\in S} \left\|C_i \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^{\infty})\right\|$. As for the remainder:

$$\begin{aligned} g^{\nu_Z-2}(t)\left\|\Delta \mathbf{w}^{(t)}\right\|^2 &= g^{\nu_Z-2}(t) \left\| \mathcal{L}_{|S}(\mathbf{w}^{(t)}) \eta_t g^{\nu_Z-1}(t) \left( \sum_{i\in S} \alpha_i^{(t)} C_i \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^{\infty}) + o(1) \right) \right\|^2 \\ &\le g^{\nu_Z-2}(t) \left(\beta^{(t)}\right)^2 \left\| \sum_{i\in S} \alpha_i^{(t)} C_i \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^{\infty}) + o(1) \right\|^2 \\ &\le g^{\nu_Z-2}(t) \left(\beta^{(t)}\right)^2 C_M^2 \\ &= o(\beta^{(t)}) \end{aligned}$$

Where the last inequality is due to $\beta^{(t)}$ decaying exponentially with $g^{(t)}$. Combining both bounds we get

$$g(t+1)^{\nu_Z} - g(t)^{\nu_Z} \le \nu_Z C_M \beta^{(t)} + o(\beta^{(t)}) \le \tilde{C}\beta^{(t)}$$

For some $\tilde{C} > 0$. Now summing from large enough $t$ to $\infty$ we get $\sum_{t=t_0}^{\infty} g(t+1)^{\nu_Z} - g(t)^{\nu_Z} = \infty$ and hence also $\sum_{t=t_0}^{\infty} \tilde{C}\beta^{(t)} = \infty$. □

Using Lemma E.4 with $\Phi(\alpha) = \frac{1}{2}\alpha^\top A \alpha = \frac{1}{2}\left\|\sum_{i\in S} \alpha_i^{(t)} C_i \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^{\infty})\right\|^2$, and Item 1 of Lemma E.5 to substitute $\varepsilon^{(t)} = o(\beta^{(t)})$, we get our update Equation (25). Hence summing up Equation (26) from $t = 0$ to $T-1$, using $\alpha^*$ some minimizer of $\Phi$, we get

$$\sum_{t=0}^{T-1} \beta^{(t)} \left(\Phi(\alpha^{(t)}) - \Phi(\alpha^*)\right) \le \mathrm{KL}\left(\alpha^* \parallel \alpha^{(0)}\right) + \sum_{t=0}^{T-1} (\beta^{(t)}\Delta^{(t)})^2/8 + 2\sum_{t=0}^{T-1} \left\|\varepsilon^{(t)}\right\|_\infty$$

Note that by the definition of $\Delta \mathbf{w}^{(t)}$ we have that $\sum_{t=0}^{T-1} \frac{(\beta^{(t)}\Delta^{(t)})^2}{8} = o(\sum_{t=0}^{T-1} \beta^{(t)})$ and by Equation (25) we have $2\sum_{t=0}^{T-1} \left\|\varepsilon^{(t)}\right\|_\infty = o(\sum_{t=0}^{T-1} \beta^{(t)})$, so dividing both sides by $\sum_{t=0}^{T-1} \beta^{(t)}$ we get

$$\frac{\sum_{t=0}^{T-1} \beta^{(t)} \left(\Phi(\alpha^{(t)}) - \Phi(\alpha^*)\right)}{\sum_{t=0}^{T-1} \beta^{(t)}} \to 0$$

Thereby $\liminf_{t\to\infty} \left(\Phi(\alpha^{(t)}) - \Phi(\alpha^*)\right) = 0$, so there exists a subsequence $t_k$ s.t. $\Phi(\alpha^{t_k}) \to \Phi(\alpha^*)$.

Notice that since $\forall i \in S$ $m_i$ has a minimal effective degree and the minimal margin among those with the minimal effective degree, hence we have that

$$(\bar{\mathbf{w}}^\infty)^\top \sum_{i\in S} \alpha_i^{(t)} C_i \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^\infty) = \sum_{i\in S} \alpha_i^{(t)} \nu_Z C_i m_i^{\nu_Z}(\bar{\mathbf{w}}^\infty) = \nu_Z \gamma_{\min}$$

In other words,

$$\sum_{i\in S} \alpha_i^{(t)} C_i \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^\infty) = (\nu_Z \gamma_{\min})\bar{\mathbf{w}}^\infty + \mathbf{v}^{(t)}$$

For $\mathbf{v}^{(t)} \perp \bar{\mathbf{w}}^\infty$ with a constant $\bar{\mathbf{w}}^\infty$ element. So minimizing $\left\|\sum_{i\in S} \alpha_i^{(t)} C_i \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^\infty)\right\|^2$ is akin to minimizing $\left\|\mathbf{v}^{(t)}\right\|^2$ for $\mathbf{v} \in \mathrm{Conv}_{i\in S}\{C_i P_\perp \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^\infty)\}$ where $P_\perp$ the projection map to the space orthogonal to $\bar{\mathbf{w}}^\infty$.

**Lemma E.6.** $\min_{\mathbf{v}\in \mathrm{Conv}_{i\in S}\{C_i P_\perp \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^\infty)\}} \|\mathbf{v}\|^2 = 0$

*Proof.* Assume towards contradiction that $0 \notin \mathrm{Conv}_{i\in S}\{C_i P_\perp \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^\infty)\}$, then by the separating hyperplane theorem there exists $\delta > 0$ and $\mathbf{q} \perp \bar{\mathbf{w}}^\infty$ (since the whole convex hull lives orthogonally to $\bar{\mathbf{w}}^\infty$) s.t. $\mathbf{q}^\top \mathbf{v} \geq \delta$ for any $\mathbf{v} \in \mathrm{Conv}_{i\in S}\{C_i P_\perp \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^\infty)\}$. Recall that $\Delta \mathbf{w}^{(t)} = \mathbf{w}^{(t+1)} - \mathbf{w}^{(t)} = \eta_t \mathcal{L}_{|S}(\mathbf{w}^{(t)}) g^{\nu_Z - 1}(t) \left(\sum_{i\in S} \alpha_i^{(t)} C_i \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^\infty) + o(1)\right)$, hence for large enough $t_0$, we have for any $t \geq t_0$ that

$$\left\langle \mathbf{q}, \Delta \mathbf{w}^{(t)} \right\rangle \geq (\delta/2)\eta_t \mathcal{L}_{|S}(\mathbf{w}^{(t)}) g^{\nu_Z - 1}(t)$$
$$\left\langle \bar{\mathbf{w}}^\infty, \Delta \mathbf{w}^{(t)} \right\rangle \leq \frac{3}{2} \nu_Z \gamma_{\min} \eta_t \mathcal{L}_{|S}(\mathbf{w}^{(t)}) g^{\nu_Z - 1}(t)$$

Summing up we get

$$\frac{\left\langle \mathbf{q}, \mathbf{w}^{(t)} - \mathbf{w}^{(t_0)} \right\rangle}{\left\langle \bar{\mathbf{w}}^\infty, \mathbf{w}^{(t)} - \mathbf{w}^{(t_0)} \right\rangle} \geq \frac{\delta \sum_{t=t_0}^{T} \eta_t \mathcal{L}_{|S}(\mathbf{w}^{(t)}) g^{\nu_Z - 1}(t)}{3\nu_Z \gamma_{\min} \sum_{t=t_0}^{T} \eta_t \mathcal{L}_{|S}(\mathbf{w}^{(t)}) g^{\nu_Z - 1}(t)} \overset{T\to\infty}{\to} \delta/3\nu_Z\gamma_{\min} > 0$$

Contradicting $\bar{\mathbf{w}}^{(t)} \to \bar{\mathbf{w}}^\infty$, since we got that $\frac{\mathbf{w}^{(t)} - \mathbf{w}^{(t_0)}}{\|\mathbf{w}^{(t)}\|} \to \bar{\mathbf{w}}^\infty$ has a nonvanishing projection on $\mathbf{q} \perp \bar{\mathbf{w}}^\infty$. □

From Lemma E.6 we have that the subsequence $t_k$ has $\sum_{i\in S} \alpha_i^{(t_k)} C_i \nabla m_i^{\nu_Z}(\bar{\mathbf{w}}^\infty) \to \nu_Z \gamma_{\min} \bar{\mathbf{w}}^\infty$. Denote the subsequence limits $\alpha_i^{(t_k)} \to \alpha_i^{t_\infty}$, and $\overline{\nabla \mathcal{L}(\mathbf{w}^{(t_\infty)})} := \lim_{k\to\infty} \frac{\nabla \mathcal{L}(\mathbf{w}^{(t_k)})}{\|\nabla \mathcal{L}(\mathbf{w}^{(t_k)})\|}$. So overall we have $\overline{\nabla \mathcal{L}(\mathbf{w}^{(t_\infty)})} \propto \sum_{i\in S} \alpha_i^{t_\infty} C_i \nabla \mathcal{P}_i^{\nu_Z}(\bar{\mathbf{w}}^\infty)^\top y_i \mathbf{x}_i \propto \bar{\mathbf{w}}^\infty$ for $\mu_i \geq 0$. We use $\alpha_i^{t_\infty}$ as the dual certificate (letting $\forall i \notin S$ $\alpha_i^{t_\infty} = 0$), that satisfies dual feasibility and complementary slackness. Therefore $\bar{\mathbf{w}}^\infty \propto \sum_{i=1}^n \alpha_i^{t_\infty} C_i \nabla \mathcal{P}_i^{\nu_i}(\bar{\mathbf{w}}^\infty)^\top \mathbf{x}_i y_i$. Scaling $\bar{\mathbf{w}}^\infty$ s.t. $\gamma_{\min} = 1$, we conclude the proof.

# F Auxiliary Material

**Theorem F.1** (Theorem 6.1 from Cai *et al.* (2025), simplified)**.** *Under Assumption 4, the expression*

$$\gamma(t) := \frac{\exp\left(\Phi\left(\exp\left(p\left(\left\|\mathbf{w}^{(t)}\right\|\right)\right)\mathcal{L}\left(\mathbf{w}^{(t)}\right)/n\right)\right)}{\left\|\mathbf{w}^{(t)}\right\|^M}$$

*Where $M = \deg(q)$ is the degree of the predictor, and $\Phi(x) = \log(\log(1/nx)) + 2/\log(nx)$, is positive and increasing from some time $t_0$ from Assumption 4.*

**Theorem F.2** (Theorem 6.2 from Cai *et al.* (2025), simplified)**.** *Under Assumption 4 the parameters* $\mathbf{w}^{(t)}$ *converge in direction to a KKT point of*

$$\min \|\mathbf{w}\|^2, \quad \forall i, \quad y_i f_H(\mathbf{w}, \mathbf{x}_i) \geq 1$$

*Where* $f_H$ *represents the top-degree homogenization of the margin (depending on* $\mathbf{x}$*).*

**Lemma F.3** (Hoeffding's Lemma)**.** *Let* $X$ *be a RV over the reals, with* $a \leq X \leq b$ *almost surely. Then* $\forall \lambda \in \mathbb{R}$

$$\mathbb{E} \exp(\lambda(X - \mathbb{E}X)) \leq \exp(\lambda^2 (b-a)^2/8)$$

**Lemma F.4.** *Let* $(e_t)_{t=1}^{\infty}, (\beta_t)_{t=1}^{\infty}, (u_t)_{t=1}^{\infty}$*, s.t.*

$$e_{t+1} = (1 - \beta_t) e_t + u_t$$

*And* $0 < \beta_t < 1$*,* $\sum_{t=1}^{\infty} \beta_t = \infty$*,* $\sum_{t=1}^{\infty} |u_t| < \infty$*. Then* $e_t \to 0$*.*

*Proof.* Iterating the recursion gives

$$e_t = e_1 \prod_{k=1}^{t-1} (1 - \beta_k) + \sum_{j=1}^{t-1} u_j \prod_{k=j+1}^{t-1} (1 - \beta_k),$$

where an empty product is understood to be equal to 1.

First,

$$0 \leq \prod_{k=1}^{t-1} (1 - \beta_k) \leq \prod_{k=1}^{t-1} e^{-\beta_k} = \exp\left( - \sum_{k=1}^{t-1} \beta_k \right) \to 0,$$

since $\sum_{k=1}^{\infty} \beta_k = \infty$. Hence

$$e_1 \prod_{k=1}^{t-1} (1 - \beta_k) \to 0.$$

It remains to treat the second term. Let $\varepsilon > 0$. Since $\sum_{j=1}^{\infty} |u_j| < \infty$, there exists $N$ such that

$$\sum_{j=N+1}^{\infty} |u_j| < \varepsilon.$$

Now split

$$\sum_{j=1}^{t-1} u_j \prod_{k=j+1}^{t-1} (1 - \beta_k) = \sum_{j=1}^{N} u_j \prod_{k=j+1}^{t-1} (1 - \beta_k) + \sum_{j=N+1}^{t-1} u_j \prod_{k=j+1}^{t-1} (1 - \beta_k).$$

For each fixed $j$, we have

$$\prod_{k=j+1}^{t-1} (1 - \beta_k) \leq \exp\left( - \sum_{k=j+1}^{t-1} \beta_k \right) \to 0,$$

so the finite sum

$$\sum_{j=1}^{N} u_j \prod_{k=j+1}^{t-1} (1 - \beta_k) \to 0.$$

For the tail, since $0 < 1 - \beta_k < 1$,

$$\left| \sum_{j=N+1}^{t-1} u_j \prod_{k=j+1}^{t-1} (1 - \beta_k) \right| \leq \sum_{j=N+1}^{t-1} |u_j| \leq \sum_{j=N+1}^{\infty} |u_j| < \varepsilon.$$

Thus the second term tends to 0. Combining this with the convergence of the first term, we conclude that

$$e_t \to 0.$$

□

# G Additional Experiments and Experimental Details

For all experiments we use Pytorch (Paszke *et al.* (2019)). We fix 20 seeds (for $i \in [4]$, $j \in [5]$, Seeds $= \{i \cdot 10^{j-1}\}$), filter experiments that did not reach interpolation or those with loss higher than $10^{-3}$, and then plot averages of 5 seeds from the remaining experiments. Each training run is 30000 epochs long. The training is run on RTX2080 GPUs, one for each experiment. Each run used an instance 8 cpus, and 12GB memory. We note that our experiments use relatively small data due to hardware memory and time constraints. We use normalized gradient descent to accelerate convergence of the dynamics, since with standard gradient descent convergence requires many more epochs, and it does not affect the implicit bias.

**Effective degree.** To estimate the effective degree of each example, we track $\min\left(\left|\frac{m_i(\mathbf{w}^{(t)})}{m_i^h(\mathbf{w}^{(t)})}\right|, \left|\frac{m_i^h(\mathbf{w}^{(t)})}{m_i(\mathbf{w}^{(t)})}\right|, \frac{\|\nabla m_i^h(\mathbf{w}^{(t)})\|}{\|\nabla m_i(\mathbf{w}^{(t)})\|}, \frac{\|\nabla m_i(\mathbf{w}^{(t)})\|}{\|\nabla m_i^h(\mathbf{w}^{(t)})\|}\right)$ for each example $\mathbf{x}_i$ and $1 \le h \le \deg(\mathcal{P}_i)$, and choose the $h$ that maximizes this quantity. Asymptotically, we expect only the effective degree to not vanish, since it goes to a constant $C$. Otherwise, at least one of the ratios should go to zero.

### G.1 Linear Autoregressive

The data $\{\mathbf{x}_i, y_i\}_{i=1}^{10}$ is random gaussian with $\mathbf{x}_i \in \mathbb{R}^{100}$ (10× overparameterization), where the labels are set by fixing a random vector of parameters $\mathbf{w}$, each coordinate sampled i.i.d. from a gaussian of scale $10^{-19}$, applying it to every example $t_i$ times, and taking the result's sign. We use a learning rate of $3 \cdot 10^{-3}$. We assign $t_i$ uniformly random from a set of indices $T \subset \mathbb{N}$. We run experiments with the $T = \{1, 2, 5\}, T = \{1, 3, 5\}, T = \{1, 4\}, T = \{2, 5\}$. Each of which we run with 20 different seeds, summing up to 80 experiments.

$T = 2, 5$

We note that the effective degree often does not equal the algebraic degree. This is significant, since our effective-degree estimator is biased toward higher degrees.

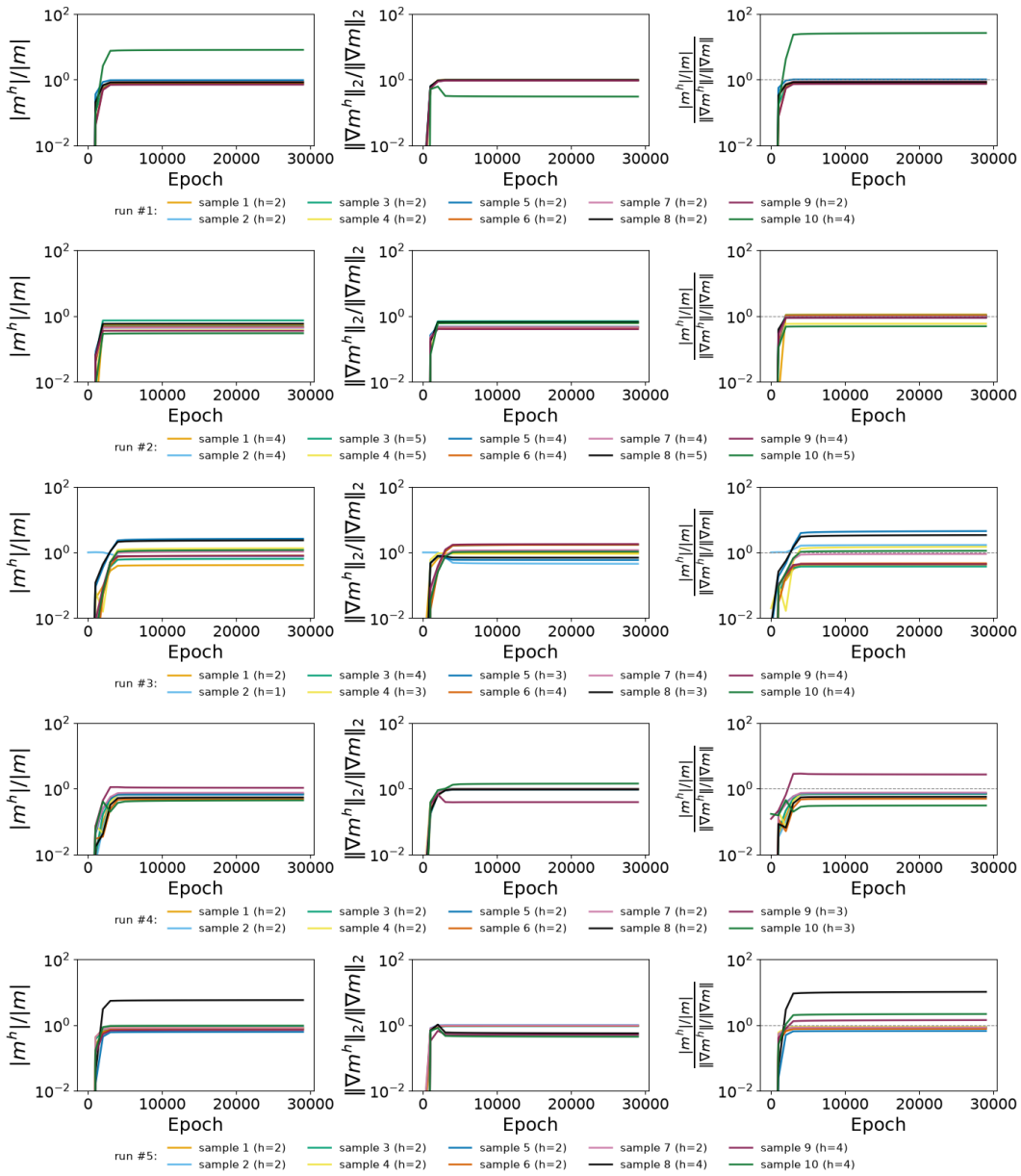


Figure 5: **Left:** Ratio between the effective-degree homogenization of the margin and the margin, $\left|m_i^h/m_i\right|$. **Middle:** The corresponding ratio for the gradients, $\left\|\nabla m_i^h\right\|/\|\nabla m_i\|$. **Right:** The ratio of the two former ratios. Each row corresponds to a different seed and shows 10 random training examples, with $h$ the effective degree of each. Y-axes are log-scale. Figure corresponding to $T = 2, 5$.

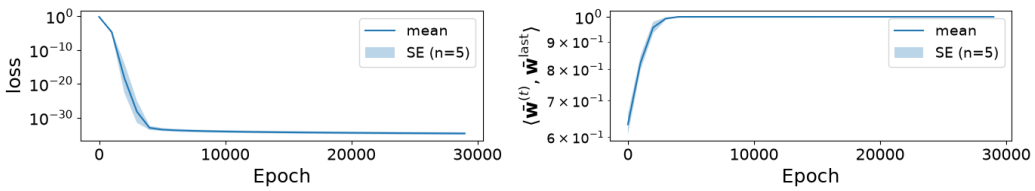


Figure 6: **Left:** The loss. **Right:** Cosine similarity to the final predictor direction. Y-axes are log-scale. Figure corresponding to $T = 2, 5$.

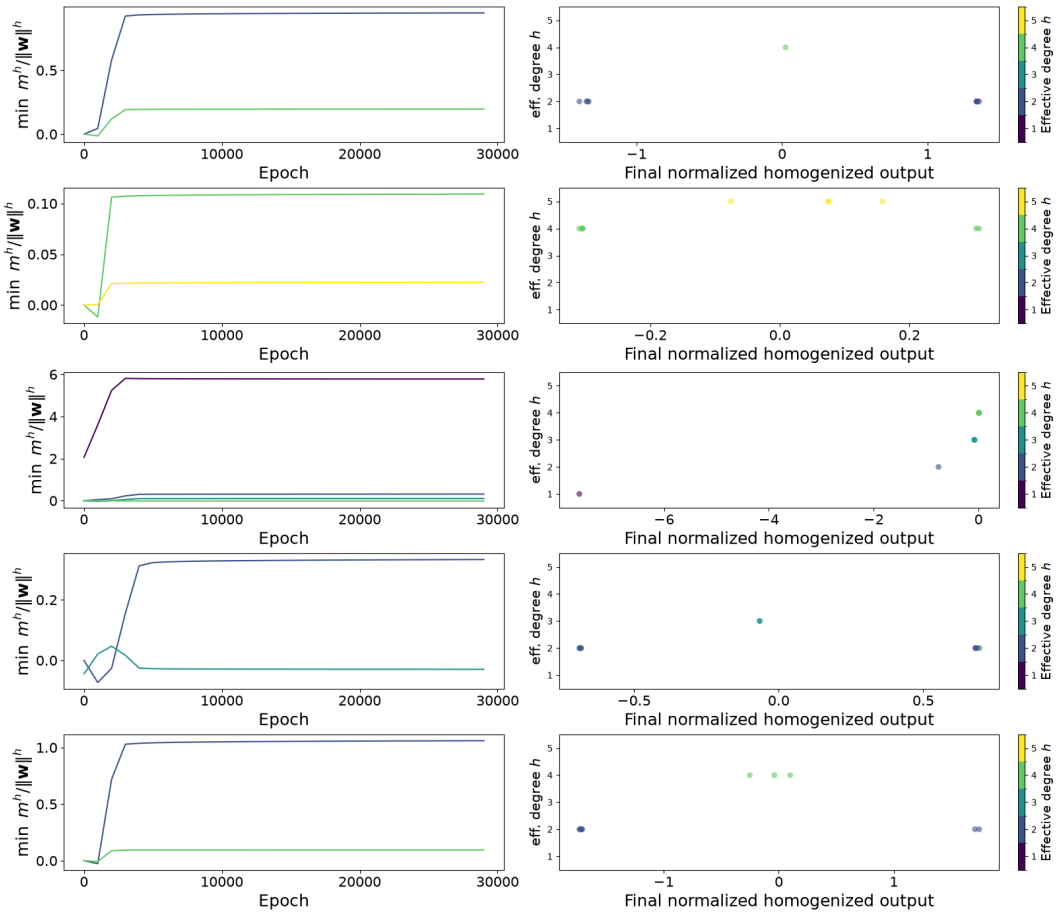


Figure 7: **Left:** Minimal normalized, homogenized margin per effective degree. **Right:** All points with their effective degrees and final normalized homogenized margins. Each row corresponds to a different seed. Figure corresponding to $T = 2, 5$.

$T = 1, 4$

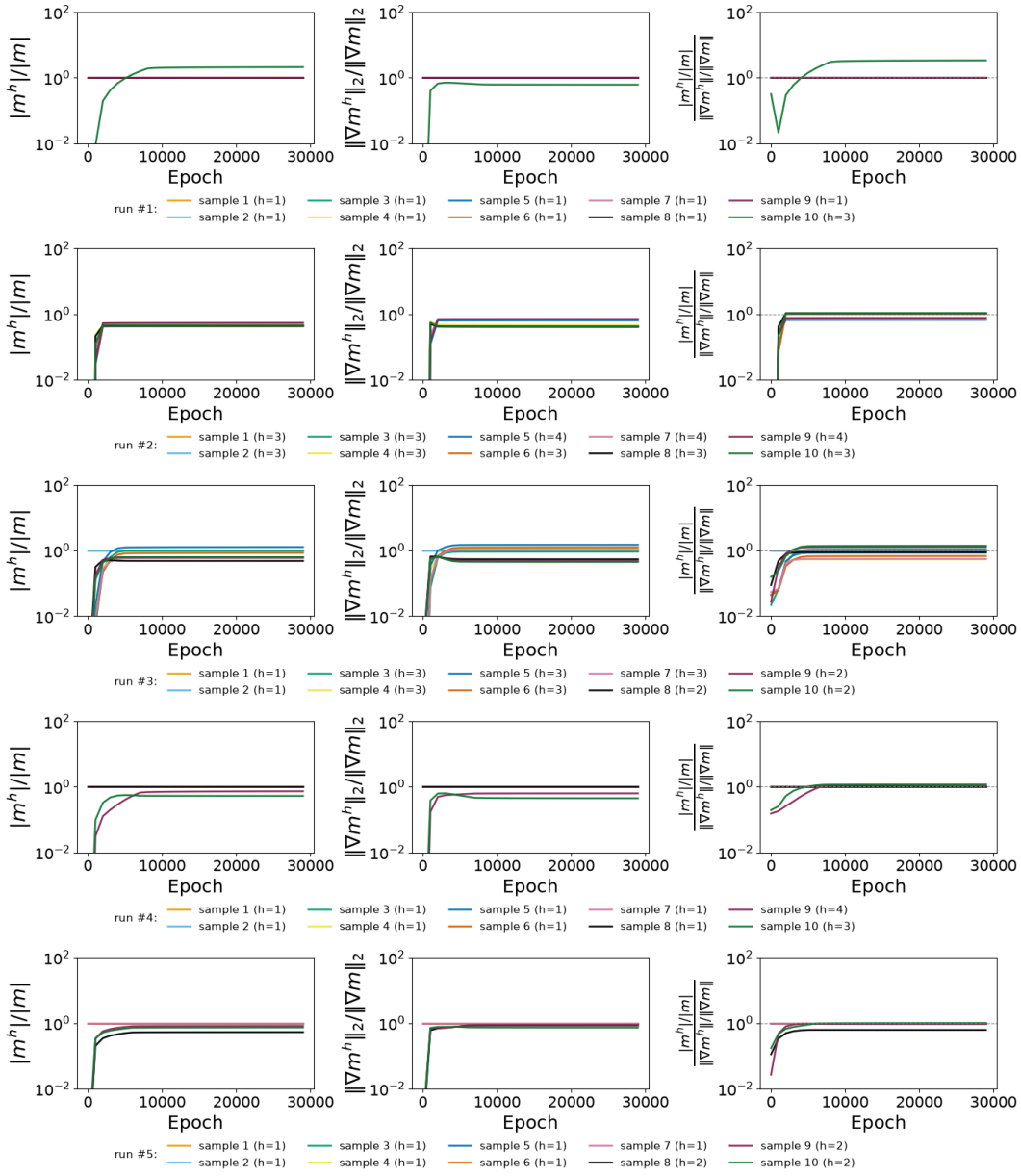


Figure 8: **Left:** Ratio between the effective-degree homogenization of the margin and the margin, $\left|m_i^h/m_i\right|$. **Middle:** The corresponding ratio for the gradients, $\left\|\nabla m_i^h\right\|/\|\nabla m_i\|$. **Right:** The ratio of the two former ratios. Each row corresponds to a different seed and shows 10 random training examples, with $h$ the effective degree of each. Y-axes are log-scale. Figure corresponding to $T = 1, 4$.

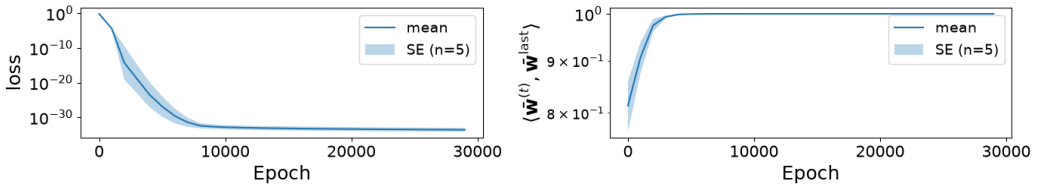


Figure 9: **Left:** The loss. **Right:** Cosine similarity to the final predictor direction. Y-axes are log-scale. Figure corresponding to $T = 1, 4$.

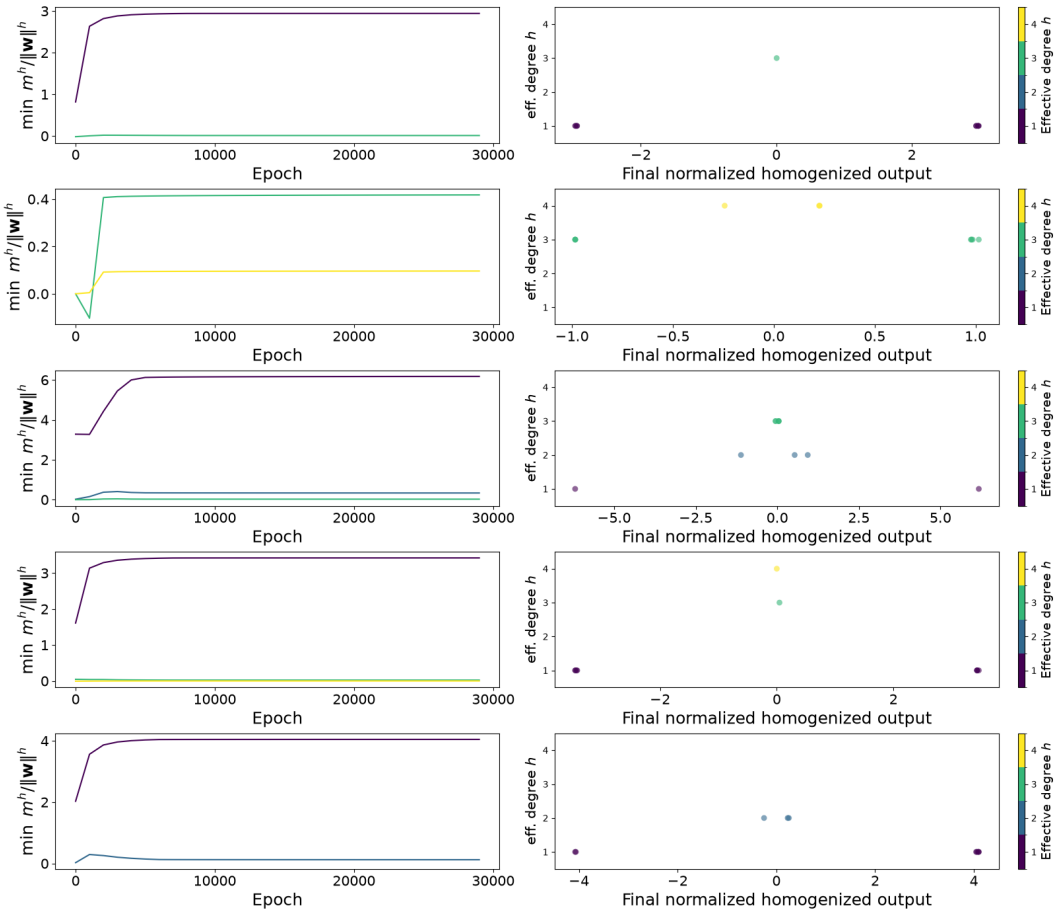


Figure 10: **Left:** Minimal normalized, homogenized margin per effective degree. **Right:** All points with their effective degrees and final normalized homogenized margins. Each row corresponds to a different seed. Figure corresponding to $T = 1, 4$.

$T = 1, 3, 5$

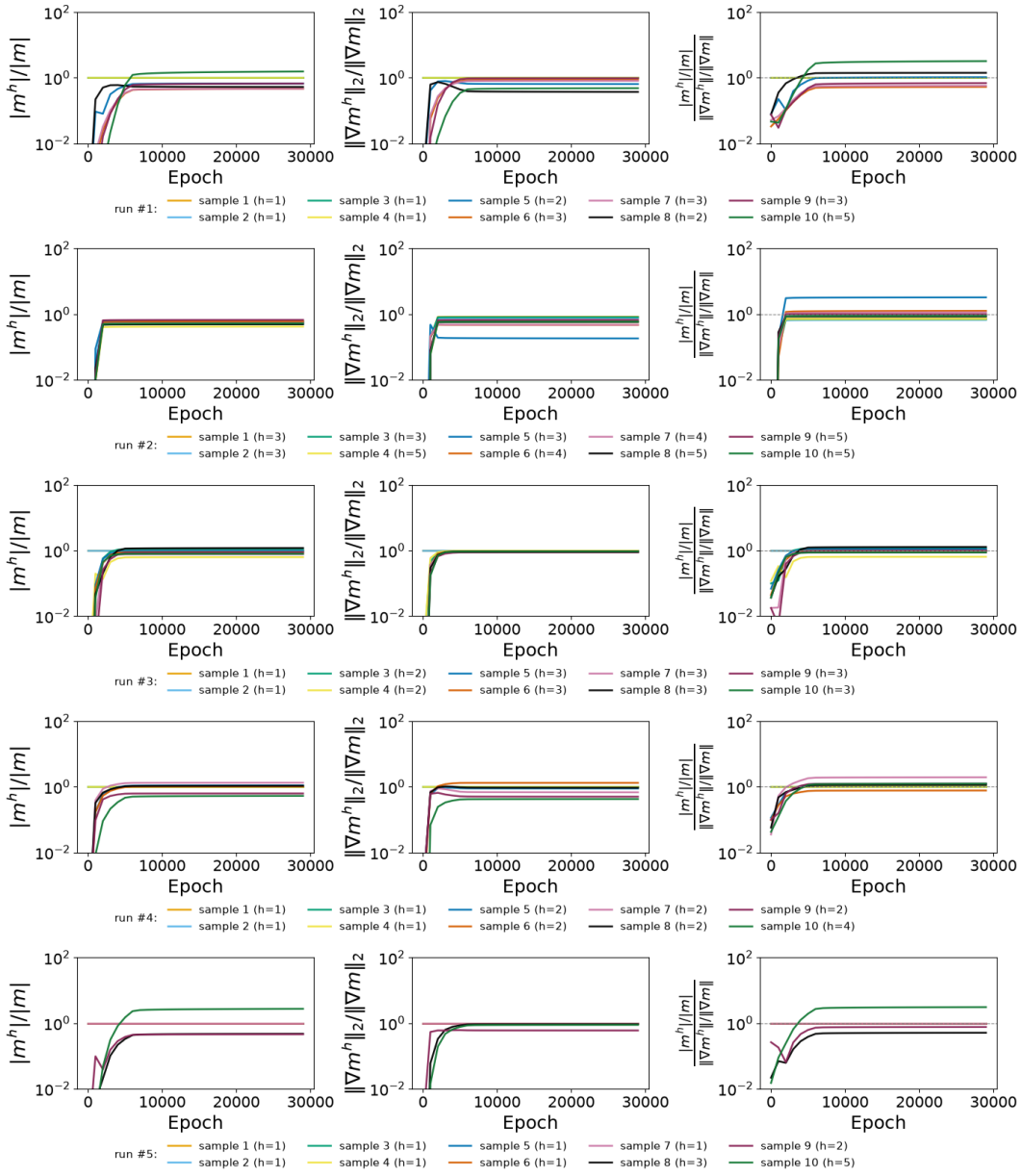


Figure 11: **Left:** Ratio between the effective-degree homogenization of the margin and the margin, $\left|m_i^h / m_i\right|$. **Middle:** The corresponding ratio for the gradients, $\left\|\nabla m_i^h\right\| / \|\nabla m_i\|$. **Right:** The ratio of the two former ratios. Each row corresponds to a different seed and shows 10 random training examples, with $h$ the effective degree of each. Y-axes are log-scale. Figure corresponding to $T = 1, 3, 5$.

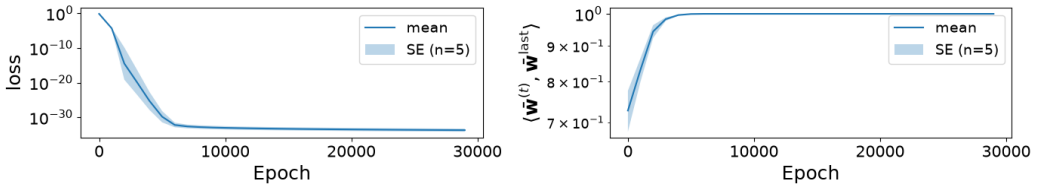


Figure 12: **Left:** The loss. **Right:** Cosine similarity to the final predictor direction. Y-axes are log-scale. Figure corresponding to $T = 1, 3, 5$.

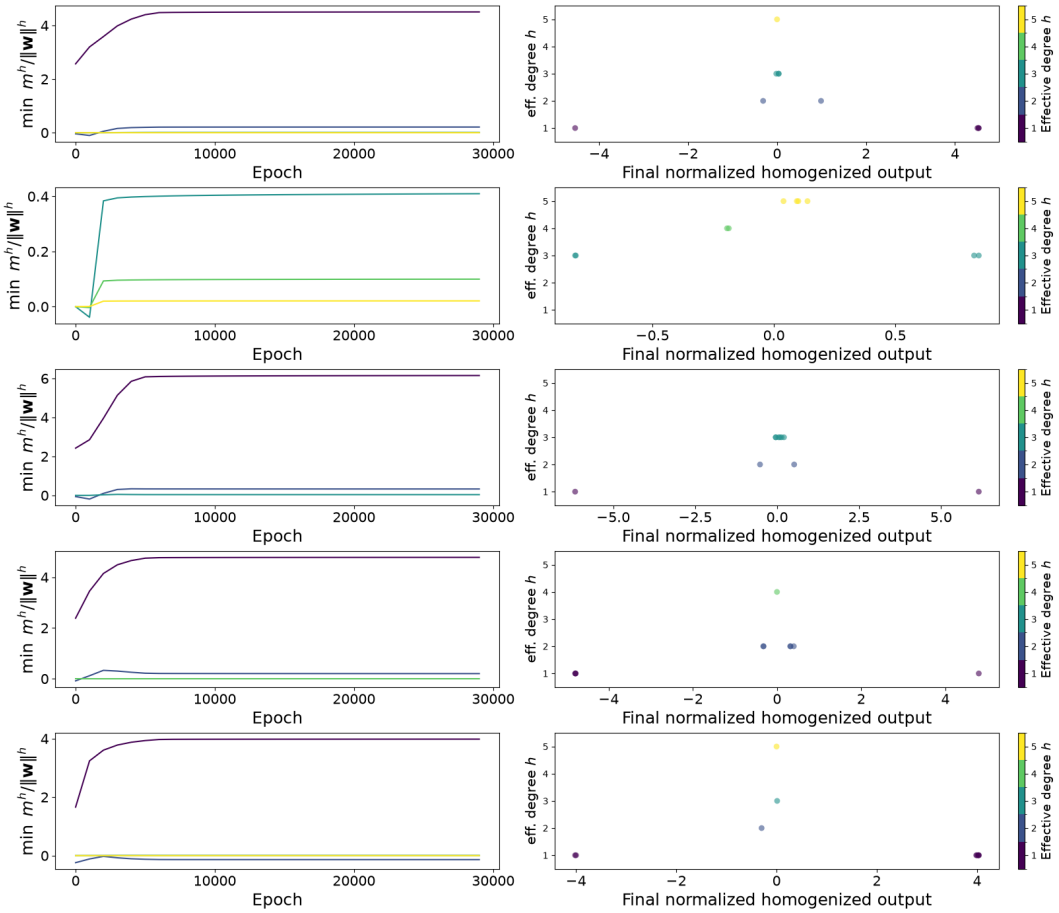


Figure 13: **Left:** Minimal normalized, homogenized margin per effective degree. **Right:** All points with their effective degrees and final normalized homogenized margins. Each row corresponds to a different seed. Figure corresponding to $T = 1, 3, 5$.

$T = 1, 2, 5$

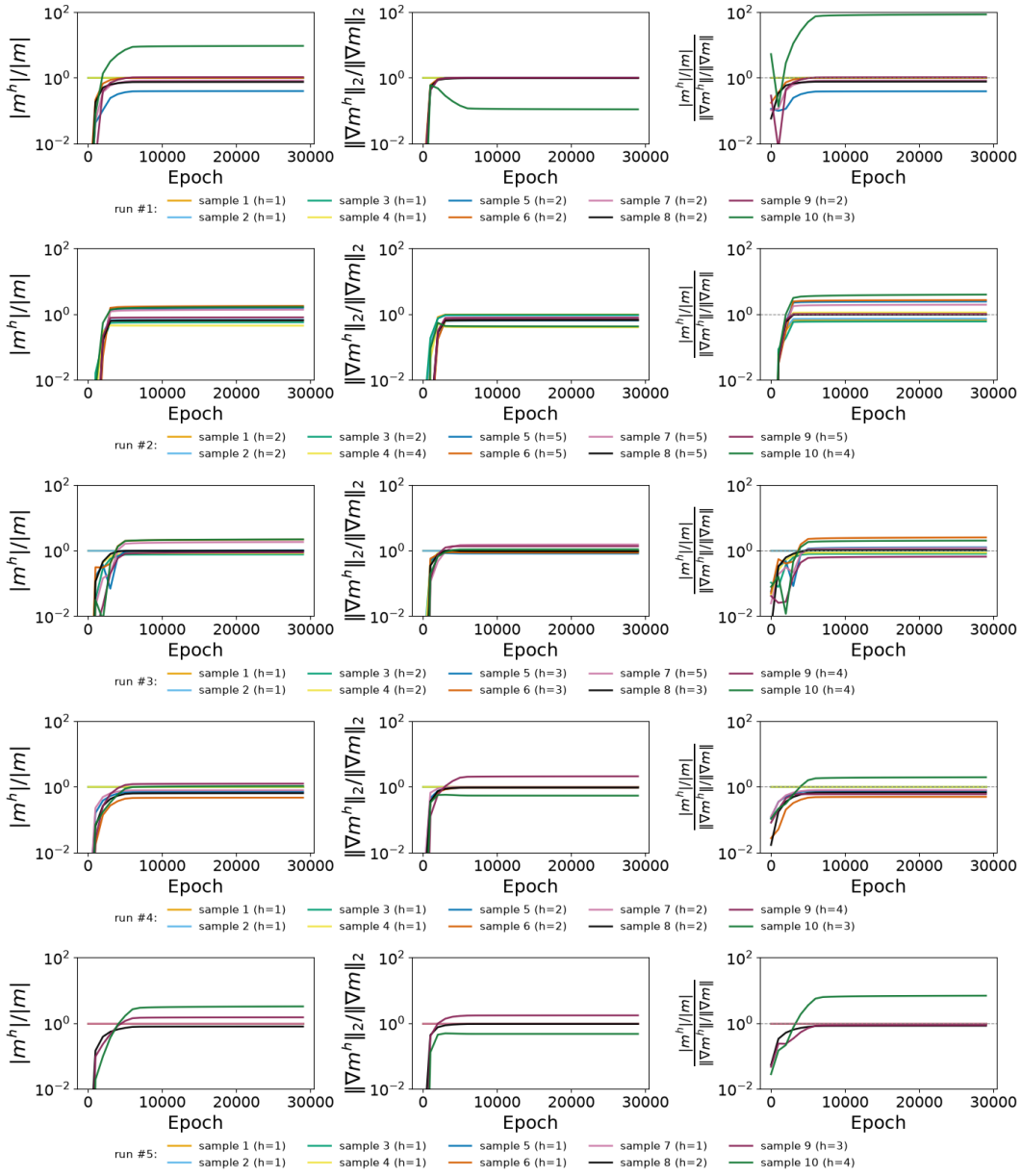


Figure 14: **Left:** Ratio between the effective-degree homogenization of the margin and the margin, $\left|m_i^h / m_i\right|$. **Middle:** The corresponding ratio for the gradients, $\left\|\nabla m_i^h\right\| / \|\nabla m_i\|$. **Right:** The ratio of the two former ratios. Each row corresponds to a different seed and shows 10 random training examples, with $h$ the effective degree of each. Y-axes are log-scale. Figure corresponding to $T = 1, 2, 5$.

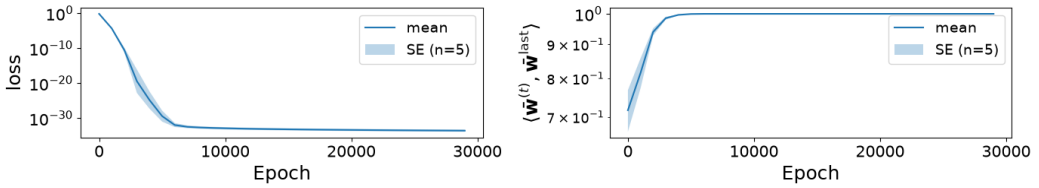


Figure 15: **Left:** The loss. **Right:** Cosine similarity to the final predictor direction. Y-axes are log-scale. Figure corresponding to $T = 1, 2, 5$.

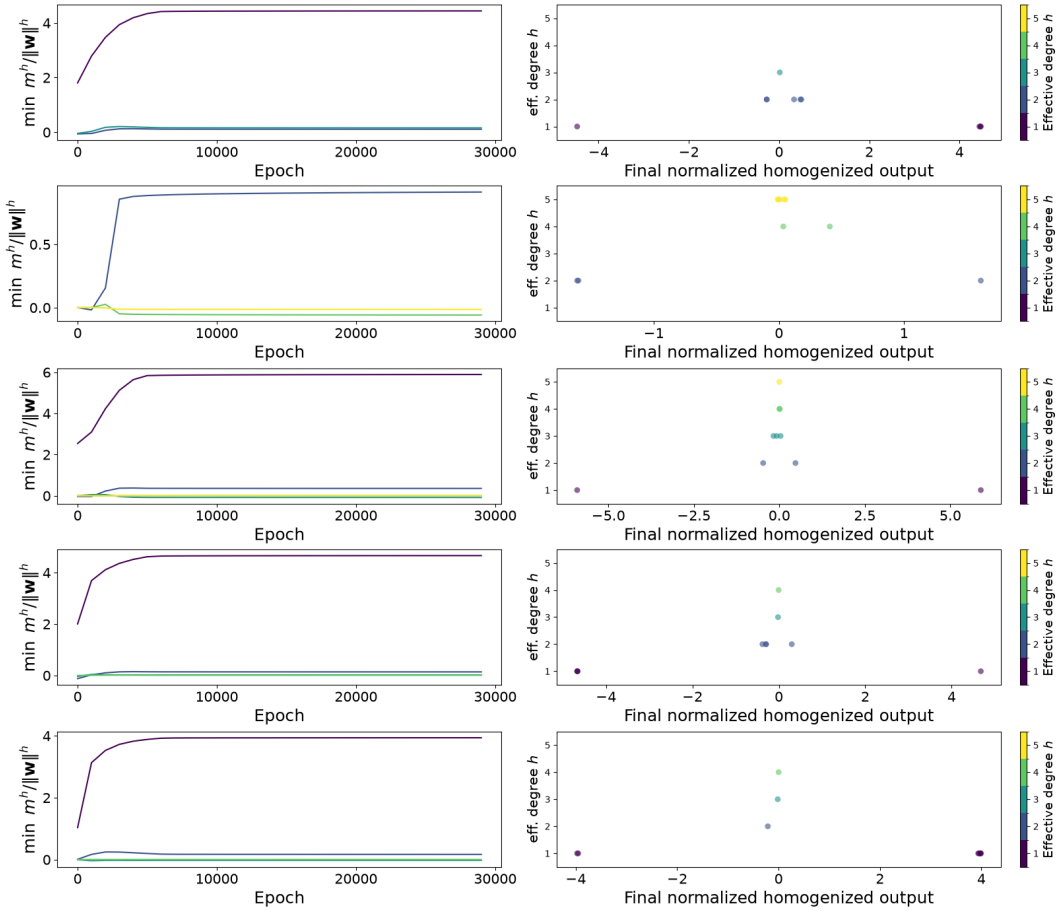


Figure 16: **Left:** Minimal normalized, homogenized margin per effective degree. **Right:** All points with their effective degrees and final normalized homogenized margins. Each row corresponds to a different seed. Figure corresponding to $T = 1, 2, 5$.

### G.2 SSM

The data $\{X_i, y_i\}_{i=1}^{10}$ is random gaussian with $X_i \in \mathbb{R}^{4\times l_i}$, where the labels are set by fixing a random projection vector $\mathbf{p} \in \mathbb{R}^4$ (which is equal to the "student's" $\mathcal{P}$ and stays fixed during training), random parameters $A \in \mathbb{R}^{100\times 100}$, $B \in \mathbb{R}^{100\times 4}$, $C \in \mathbb{R}^{100\times 4}$, each coordinate sampled i.i.d. from a gaussian of scale $10^{-19}$, applying it to every example $t_i$ times, and taking the result's sign. We assign $l_i, t_i$ uniformly random from a set of indices $T \subset \mathbb{N} \times \mathbb{N}$. We run experiments with the sets $T = \{(3, 1), (3, 2), (5, 2)\}, T = \{(1, 1); (2, 2); (3, 3)\}, T = \{(3, 1), (2, 2), (1, 3)\}$. Each of which we run with 20 different seeds, summing up to 80 experiments.

$l, T = (3,1), (3,2), (5,2)$

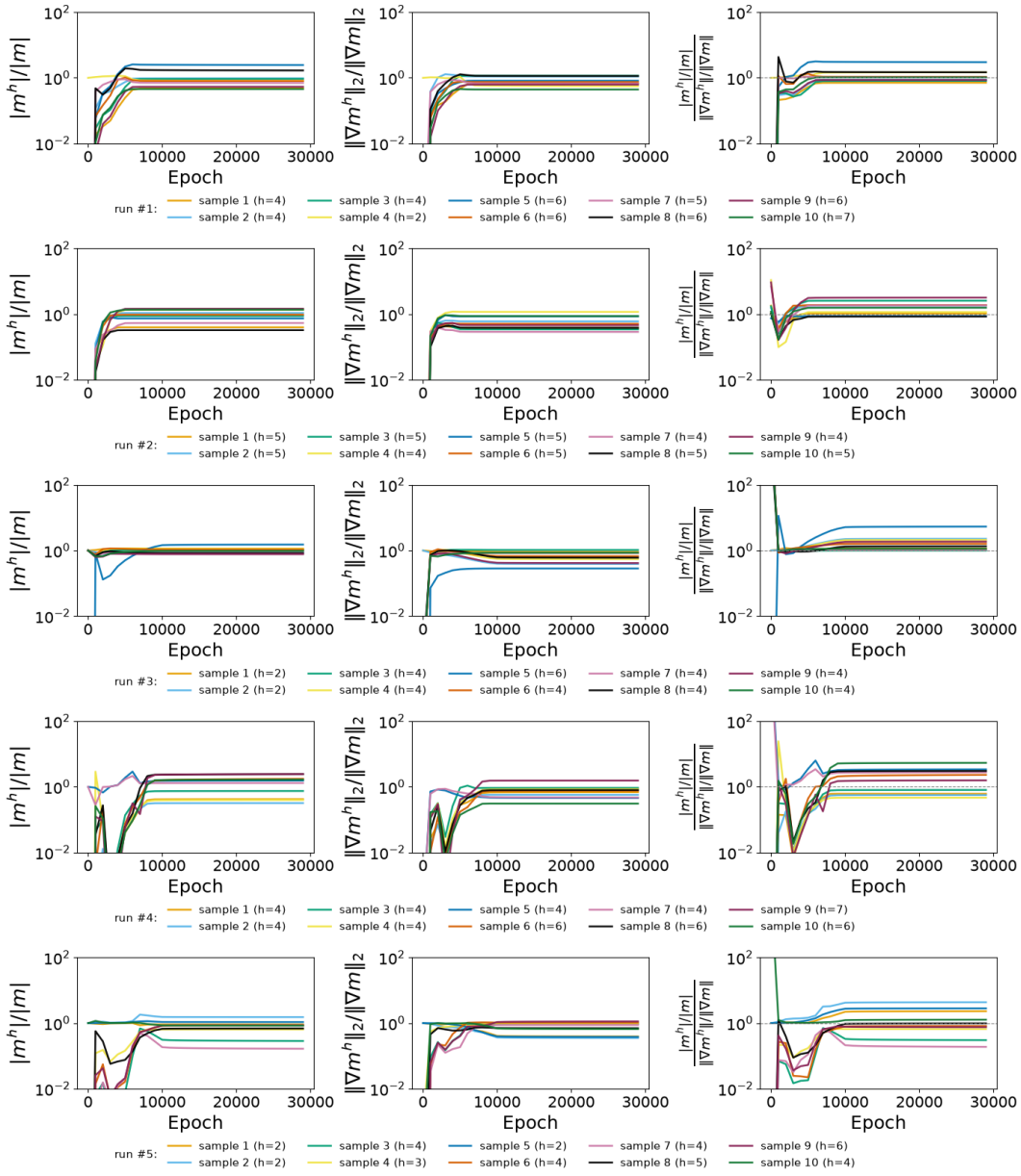


Figure 17: **Left:** Ratio between the effective-degree homogenization of the margin and the margin, $\left|m_i^h / m_i\right|$. **Middle:** The corresponding ratio for the gradients, $\left\|\nabla m_i^h\right\| / \|\nabla m_i\|$. **Right:** The ratio of the two former ratios. Each row corresponds to a different seed and shows 10 random training examples, with $h$ the effective degree of each. Y-axes are log-scale. The figure is for $l, T = (3,1), (3,2), (5,2)$.

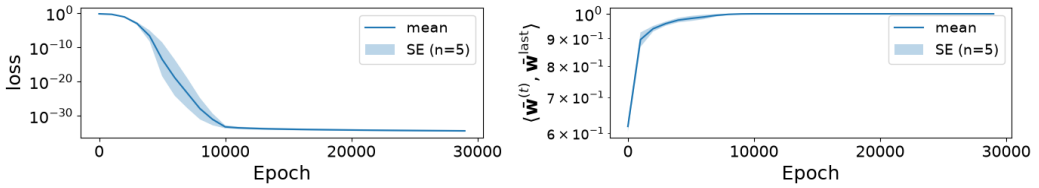


Figure 18: **Left:** The loss. **Right:** Cosine similarity to the final predictor direction. Y-axes are log-scale. The figure is for $l, T = (3, 1), (3, 2), (5, 2)$.

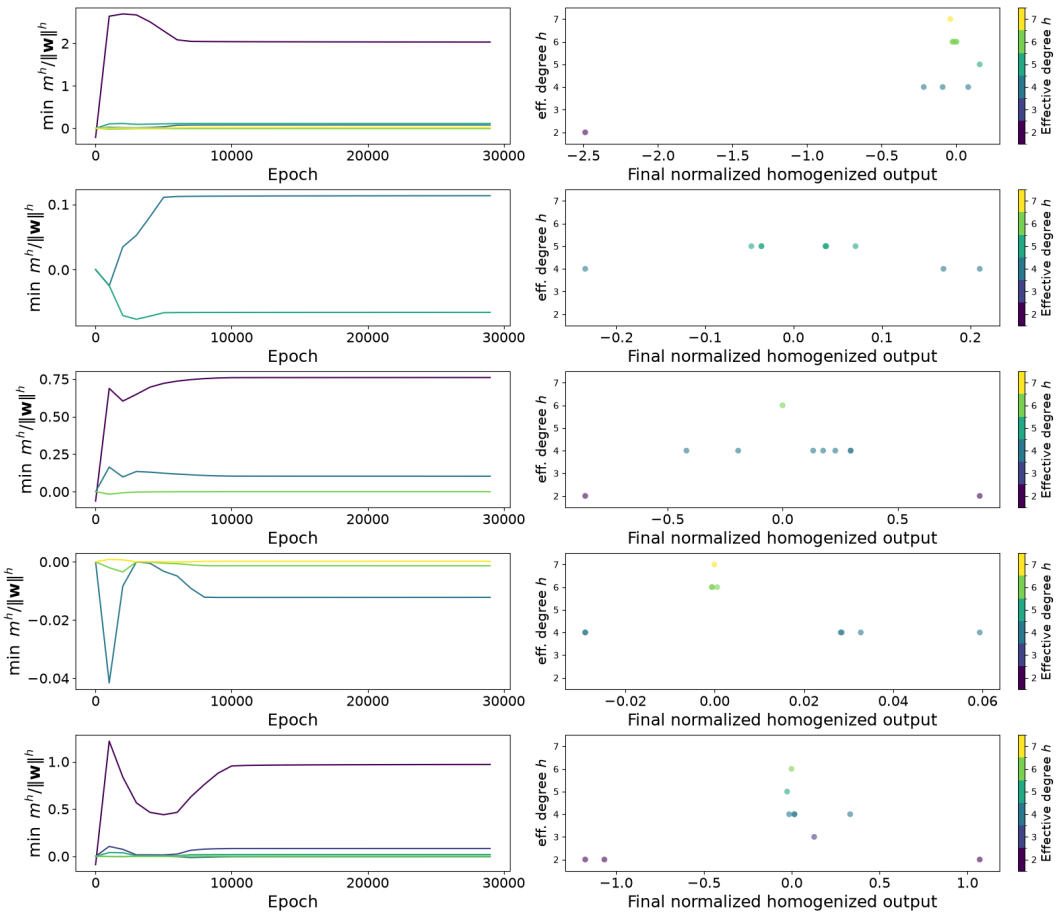


Figure 19: **Left:** Minimal normalized, homogenized margin per effective degree. **Right:** All points with their effective degrees and final normalized homogenized margins. Each row corresponds to a different seed. The figure is for $l, T = (3, 1), (3, 2), (5, 2)$.

$l, T = (1,1), (2,2), (3,3)$

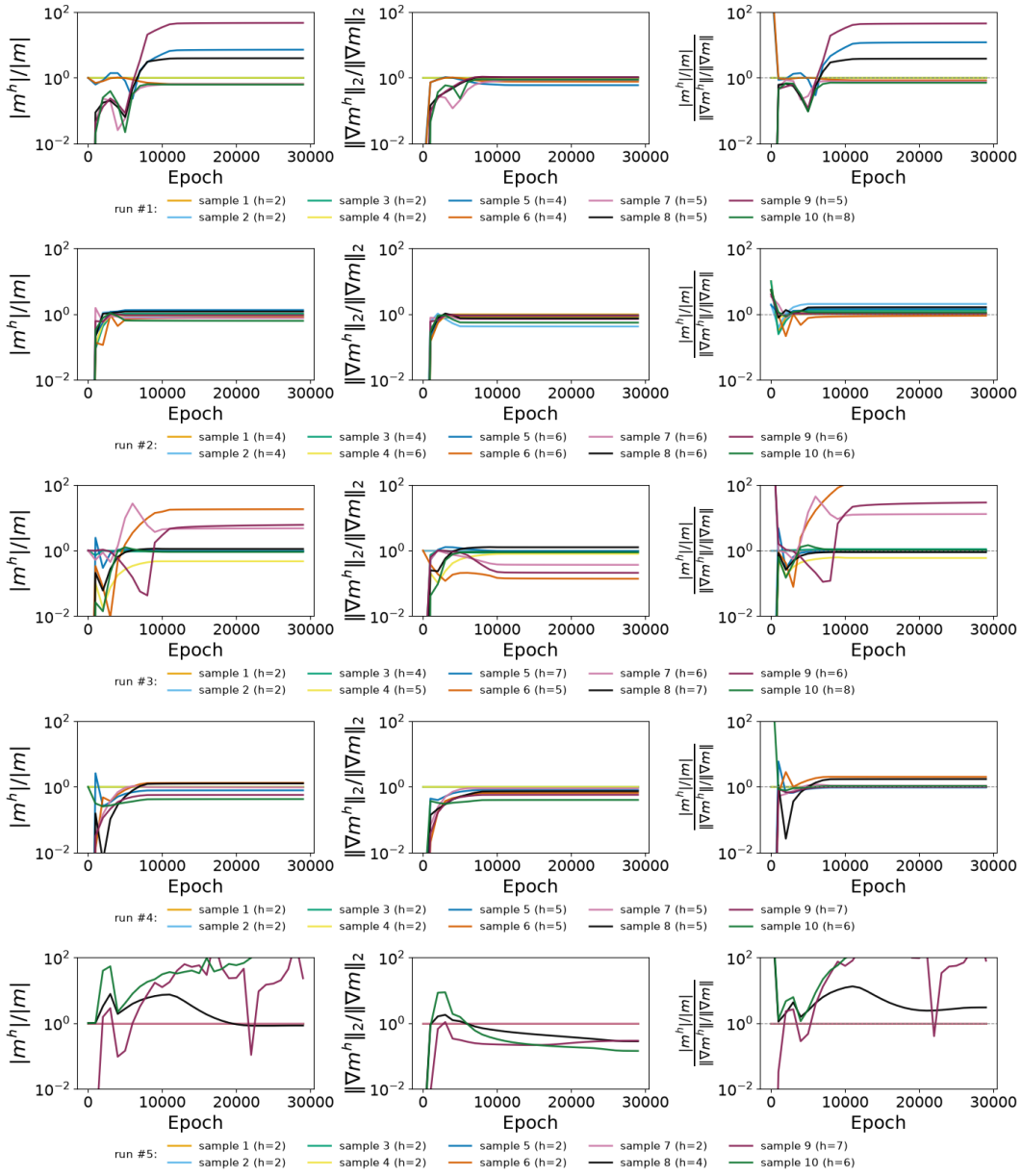


Figure 20: **Left:** Ratio between the effective-degree homogenization of the margin and the margin, $\left|m_i^h/m_i\right|$. **Middle:** The corresponding ratio for the gradients, $\left\|\nabla m_i^h\right\|/\|\nabla m_i\|$. **Right:** The ratio of the two former ratios. Each row corresponds to a different seed and shows 10 random training examples, with $h$ the effective degree of each. Y-axes are log-scale. The figure is for $l, T = (1,1), (2,2), (3,3)$.

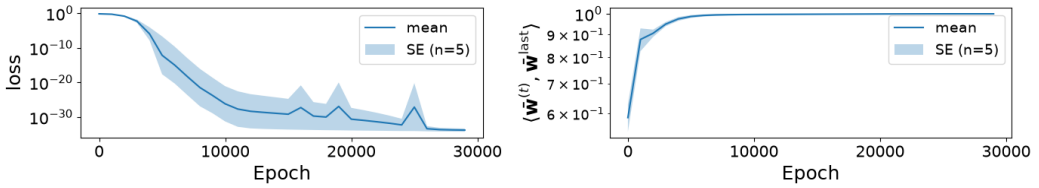


Figure 21: **Left:** The loss. **Right:** Cosine similarity to the final predictor direction. Y-axes are log-scale. The figure is for $l, T = (1, 1), (2, 2), (3, 3)$.

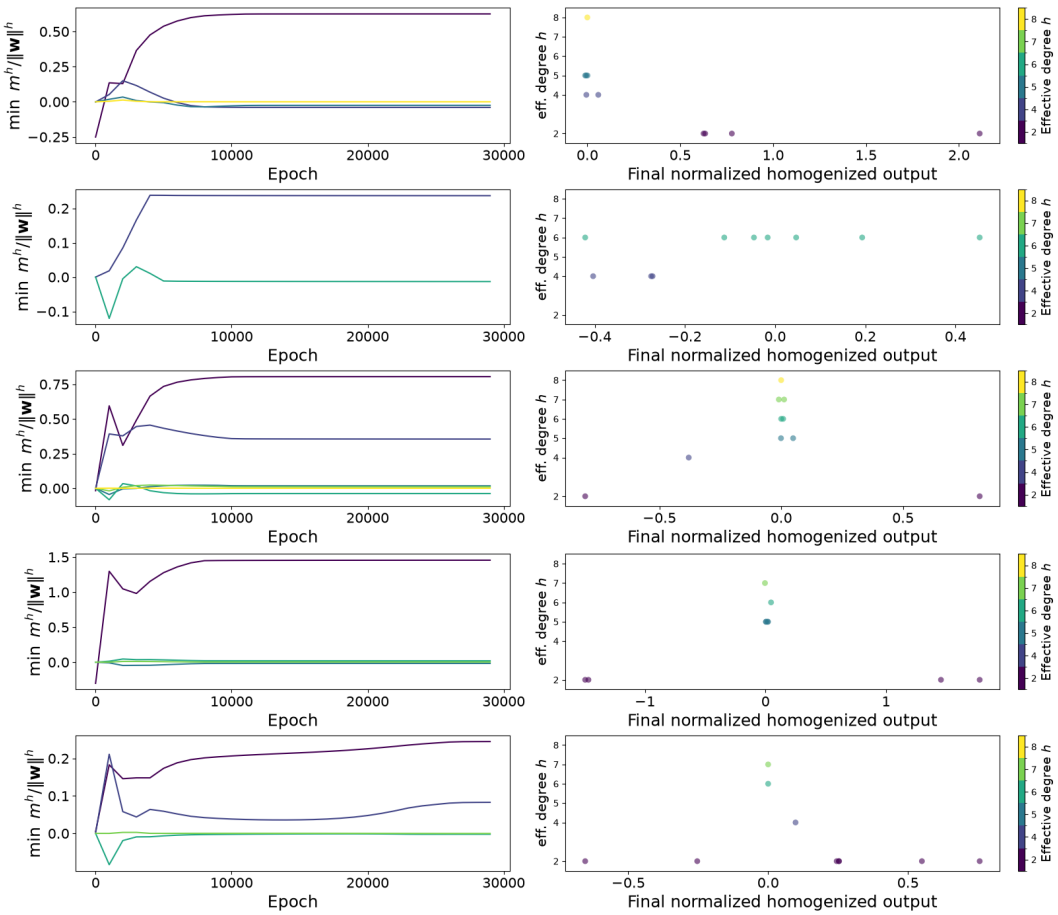


Figure 22: **Left:** Minimal normalized, homogenized margin per effective degree. **Right:** All points with their effective degrees and final normalized homogenized margins. Each row corresponds to a different seed. The figure is for $l, T = (1, 1), (2, 2), (3, 3)$.

$l, T = (3,1), (2,2), (1,3)$

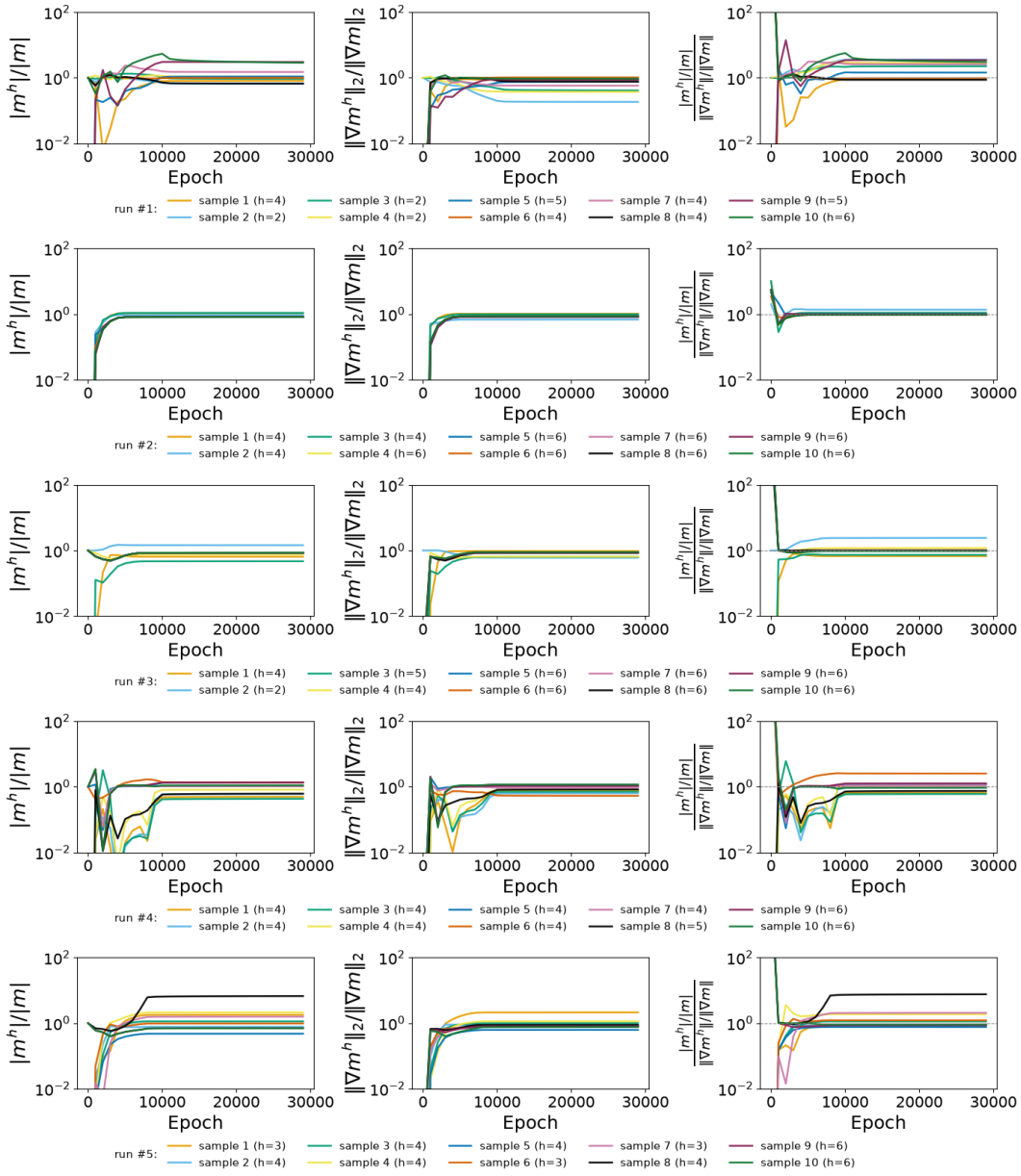


Figure 23: **Left:** Ratio between the effective-degree homogenization of the margin and the margin, $\left|m_i^h / m_i\right|$. **Middle:** The corresponding ratio for the gradients, $\left\|\nabla m_i^h\right\| / \|\nabla m_i\|$. **Right:** The ratio of the two former ratios. Each row corresponds to a different seed and shows 10 random training examples, with $h$ the effective degree of each. Y-axes are log-scale. The figure is for $l, T = (3,1), (2,2), (1,3)$.

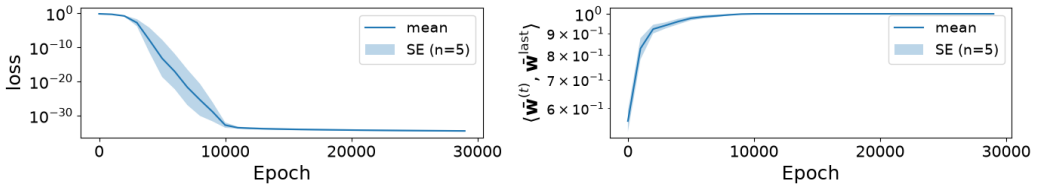


Figure 24: **Left:** The loss. **Right:** Cosine similarity to the final predictor direction. Y-axes are log-scale. The figure is for $l, T = (3,1), (2,2), (1,3)$.

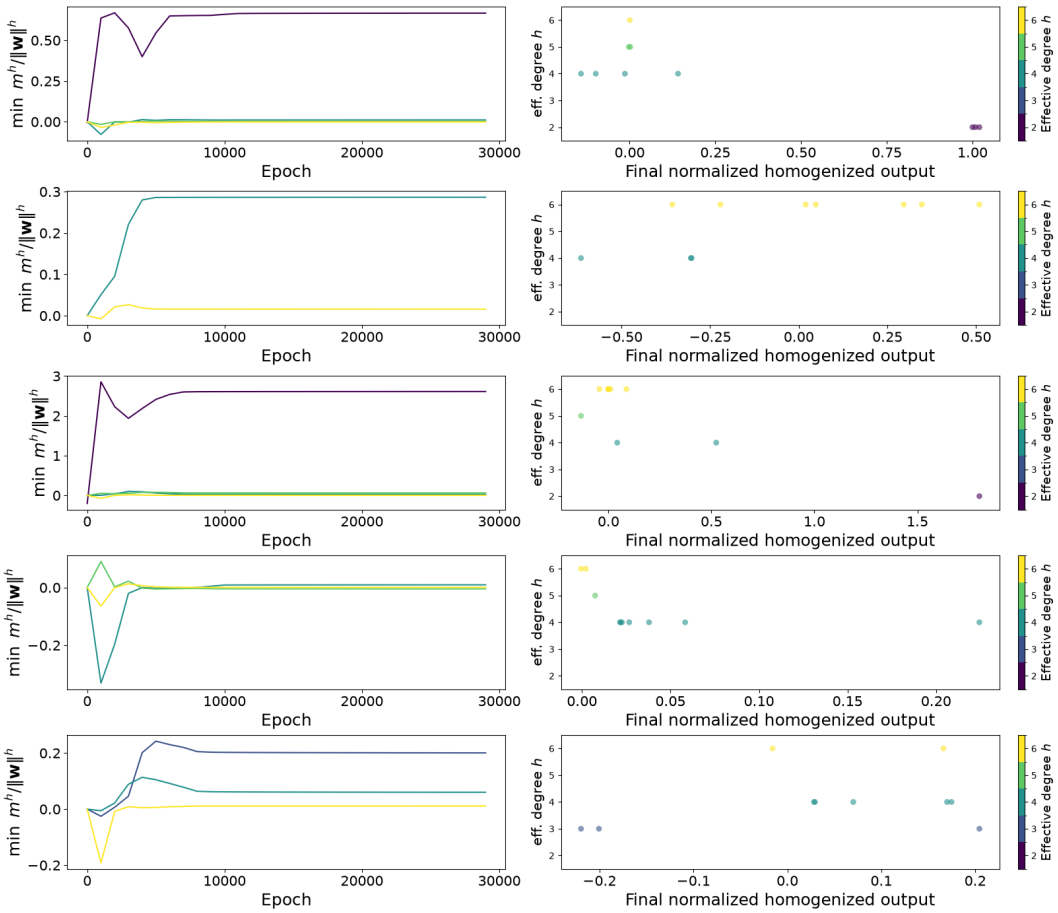


Figure 25: **Left:** Minimal normalized, homogenized margin per effective degree. **Right:** All points with their effective degrees and final normalized homogenized margins. Each row corresponds to a different seed. The figure is for $l, T = (3,1), (2,2), (1,3)$.

### G.3 MNIST SSM

We use data from MNIST (LeCun, 1998). The data $\{X_i, y_i\}_{i=1}^{10}$ divided to be half of images of the number 8 (and positive label), and half of the number 1 (and negative label). $X_i \in \mathbb{R}^{28 \times l_i}$, $l_i$ is chosen uniformly at random from $\{14, 21, 28\}$, and we truncate the rows up to that point (data is first transposed so dimensions would fit). That is, we refer to the image's rows as a sequence, each row is a token of dimension 28, and some examples dont use all the rows, in order for the result to have diverse effective degrees. Random parameters $A \in \mathbb{R}^{64 \times 64}$, $B \in \mathbb{R}^{64 \times 28}$, $C \in \mathbb{R}^{64 \times 28}$, $\mathbf{p} \in \mathbb{R}^{28}$, each coordinate sampled i.i.d. from a gaussian of scale $10^{-19}$. We run the experiment with 20 different seeds with learning rate $10^{-3}$.

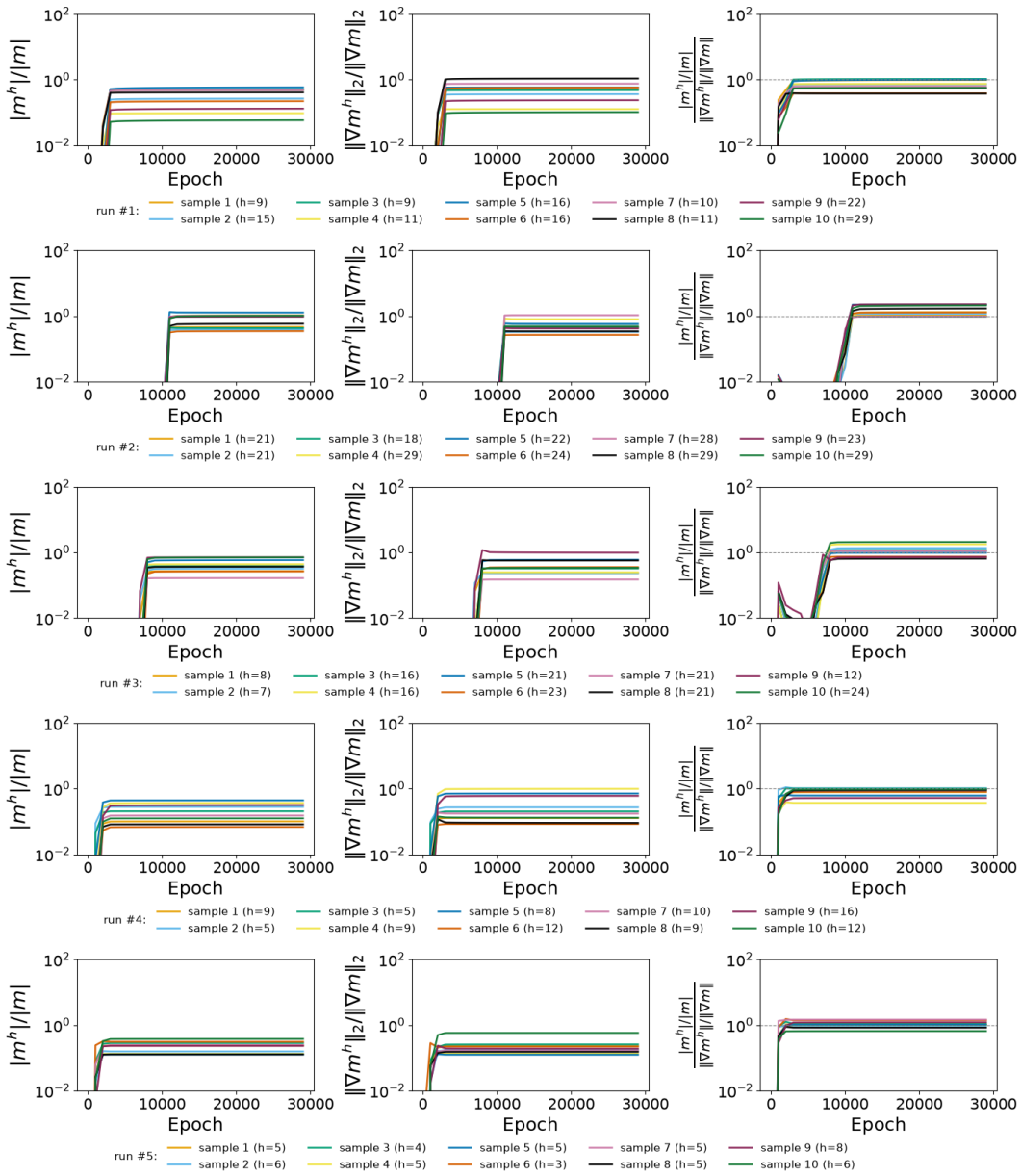


Figure 26: **Left:** Ratio between the effective-degree homogenization of the margin and the margin, $\left|m_i^h/m_i\right|$. **Middle:** The corresponding ratio for the gradients, $\left\|\nabla m_i^h\right\|/\|\nabla m_i\|$. **Right:** The ratio of the two former ratios. Each row corresponds to a different seed and shows 10 random training examples, with $h$ the effective degree of each. Y-axes are log-scale. The figure is for SSM trained on MNIST with $l = \{14, 21, 28\}$.

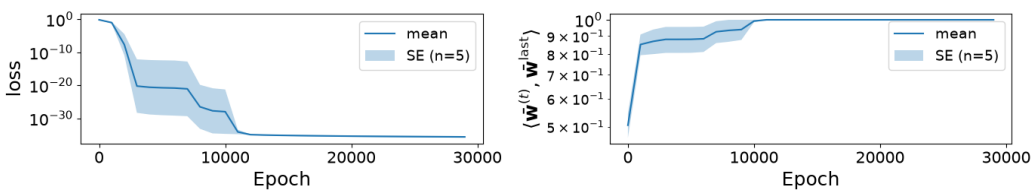


Figure 27: **Left:** The loss. **Right:** Cosine similarity to the final predictor direction. Y-axes are log-scale. The figure is for SSM trained on MNIST with $l = \{14, 21, 28\}$.

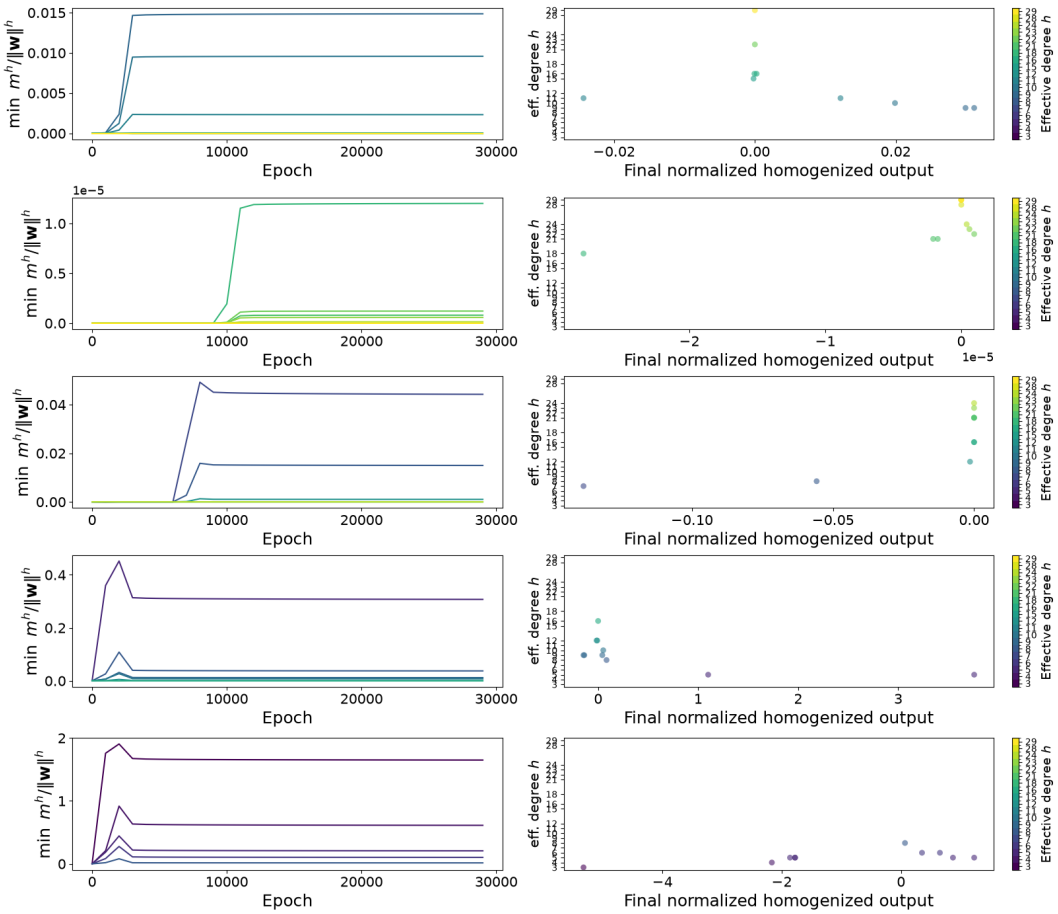


Figure 28: **Left:** Minimal normalized, homogenized margin per effective degree. **Right:** All points with their effective degrees and final normalized homogenized margins. Each row corresponds to a different seed. The figure is for SSM trained on MNIST with $l = \{14, 21, 28\}$.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We repeat the claims in those sections in our Results section (Section 5) showing proofs in Appendix E

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: While we don't have a specific limitations section, throughout the paper we discuss the various assumptions we make, and show that in some cases they don't hold Lemma A.1

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: See our results section Section 5 and our proofs (Appendix E). Other lemmas are in Section 4.2 and proofs are in Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See full details in the appendix appendix G

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: we provide a README with the code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We specify everything in Appendix G

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All relevant figures have error bars. E.g. Figure 2

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appears in Appendix G

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We read the guidelines and made sure our paper complies with them.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our work is mainly theoretical and provides explanations for existing practical methods rather than suggesting new practical methods that may have societal effect.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The experiments are done with small models, mostly synthetic data, which poses no risk of releasing.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We credit MNIST and Pytorch in Appendix G

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code is well documented.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The only humans involved in this research are the researchers.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No study participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Usage for editing purposes only.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.