TOWARDS UNDERSTANDING HOW MOMENTUM IMPROVES GENERALIZATION IN DEEP LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic gradient descent (SGD) with momentum is widely used for training modern deep learning architectures. While it is well understood that using momentum can lead to faster convergence rate in various settings, it has also been observed that momentum yields higher generalization. Prior work argue that momentum stabilizes the SGD noise during training and this leads to higher generalization. In this paper, we take the opposite view to this result and first empirically show that gradient descent with momentum (GD+M) significantly improves generalization comparing to gradient descent (GD) in many deep learning tasks. From this observation, we formally study how momentum improves generalization in deep learning. We devise a binary classification setting where a two-layer (over-parameterized) convolutional neural network trained with GD+M provably generalizes better than the same network trained with vanilla GD, when both algorithms start from the same random initialization. The key insight in our analysis is that momentum is beneficial in datasets where the examples share some features but differ in their margin. Contrary to the GD model that memorizes the small margin data, GD+M can still learn the features in these data thanks to its historical gradients. We also empirically verify this learning process of momentum in real-world settings.

1 INTRODUCTION

It is commonly accepted that adding momentum to an optimization algorithm is required to optimally train a large-scale deep network. Most of the modern architectures maintain during the training process a heavy momentum close to 1 (Krizhevsky et al., 2012; Simonyan & Zisserman, 2014; He et al., 2016; Zagoruyko & Komodakis, 2016). Indeed, it has been empirically observed that architectures trained with momentum outperform those which are trained without (Sutskever et al., 2013). Several papers have attempted to explain this phenomenon. From the optimization perspective, Defazio (2020) assert that momentum yields faster convergence of the training loss since, at the early stages, it cancels out the noise from the stochastic gradients. On the other hand, Leclerc & Madry (2020) empirically observes that momentum yields faster training convergence only when the learning rate is small. While these works shed light on how momentum acts on neural network training, they fail to capture the generalization improvement induced by momentum (Sutskever et al., 2013). Besides, the noise reduction property of momentum advocated by Defazio (2020) seems to even contradict the observation that, in deep learning, having a large noise in the training improves generalization (Li et al., 2019; HaoChen et al., 2020). To the best of our knowledge, there is no existing work which *theoretically explains* how momentum improves generalization in deep learning. Therefore, this paper aims to close this gap and addresses the following question:

Is the higher generalization induced by momentum tied to the stochastic noise of the gradient? If not, what is the underlying mechanism of momentum improving generalization in deep learning?

In this paper, we empirically verify that the generalization improvement induced by momentum *is not* tied to the stochasticity of the gradient. Indeed, as reported in Figure 1, momentum improves generalization more significantly for full batch GD than for SGD in CIFAR object recognition tasks. Motivated by this empirical observation and the fact that the stochastic noise influences generalization, we theoretically study how gradient descent with momentum (GD+M) can generalize better than vanilla gradient descent (GD). We therefore *only focus on the contribution of momentum of the true gradient on generalization*.

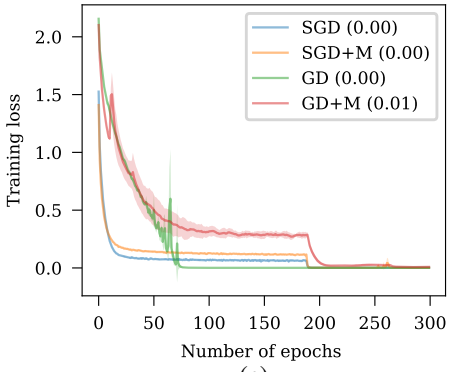
The question we address concerns algorithmic regularization which characterizes the generalization of an optimization algorithm when multiple global solutions exist in over-parameterized deep learning

	CIFAR-10		CIFAR-100	
	Test	Ratio	Test	Ratio
R18	75.83/84.68	1.11	43.32/51.99	1.20
WR16	75.02/84.48	1.12	42.95/51.33	1.20

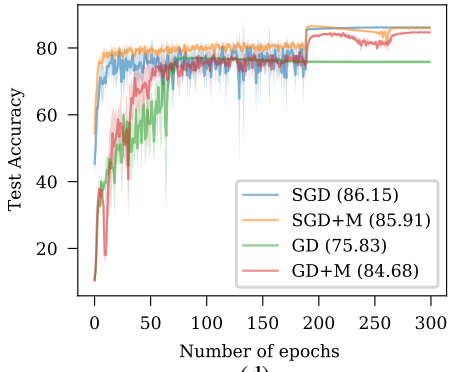
(a)

	CIFAR-10		CIFAR-100	
	Test	Ratio	Test	Ratio
R18	86.15/85.91	0.99	53.81/58.01	1.08
WR16	84.83/87.85	1.04	55.09/60.83	1.10

(b)



(c)



(d)

Figure 1: Test accuracy obtained with Resnet-18 (R18) and WideResnet16 (WR16) on CIFAR-10 and CIFAR-100. The architectures are trained using GD/GD+M (a) and SGD/ SGD+M (b) for 300 epochs to ensure zero training error. (c)-(d) respectively display the training loss and test accuracy by R18 with GD/GD+M on CIFAR-10. To isolate the effect of momentum, we *turn off* data augmentation, dropout and batch normalization. GD and SGD respectively refer to stochastic gradient descent with batch sizes $50k$ (full batch) and 128. We grid searched the best (scheduled) learning rate and weight decay for each individual algorithm separately. Results are averaged over 3 runs and we only report the mean (see Appendix for complete table).

model Soudry et al. (2018); Lyu & Li (2019); Ji & Telgarsky (2019); Chizat & Bach (2020); Gunasekar et al. (2018); Arora et al. (2019). This regularization arises in deep learning mainly due to the *non-convexity* of the objective function. Indeed, this latter can create multiple global minima scattered in the space that vastly differ in terms of generalization. Algorithmic regularization is induced by and depends on many factors such as learning rate and batch size (Goyal et al., 2017; Hoffer et al., 2017; Keskar et al., 2016; Smith et al., 2018), initialization Allen-Zhu & Li (2020), adaptive step-size (Kingma & Ba, 2014; Neyshabur et al., 2015; Wilson et al., 2017), batch normalization (Arora et al., 2018; Hoffer et al., 2019; Ioffe & Szegedy, 2015) and dropout (Srivastava et al., 2014; Wei et al., 2020). However, none of these works theoretically analyzes the regularization induced by momentum. We therefore start our investigation by raising the following question:

Does momentum ***unconditionally*** *improve generalization in deep learning?*

This question could be positively answered given the success of momentum for learning distinct architectures such as ResNets (He et al., 2016) or BERT (Devlin et al., 2018). However, we here empirically give a negative answer through the following synthetic example in deep learning. We consider a binary classification problem where data-points are generated from a standard normal distribution and labels are outputs of teacher networks. Starting from the same initialization, we train different over-parametrized student networks using GD and GD+M. Based on Table 1, whether the target function is simple (linear) or complex (neural network), momentum does not improve generalization even when using a non-linear neural network as learner. The same observation holds for SGD/SGD+M as shown in the Appendix. Therefore, momentum *does not* always lead to a higher generalization in deep learning. Instead, such benefit seems to heavily depend on both the *structure of the data* and the *learning problem*.

On which data set does momentum help generalization? In this paper, **in order to determine the underlying mechanism produced by momentum to improve generalization**, we design a binary classification problem with a simple data structure where training a two-layer (over-parameterized) convolutional network with momentum provably improves generalization in deep learning. It is built upon a data distribution that relies on the concepts of *feature* and *margin*. Informally, each example in this distribution is a 1D image having P patches. One of the patches (the signal patch) contains a feature we want to learn and all the others are Gaussian random noise with small variance.

Mathematically, one can think of a feature as a vector $w^* \in \mathbb{R}^d$. We assume that our training examples are divided into *large margin* data where the signal is αw^* with α constant and *small margin* data where the signal is βw^* with $\beta \ll 1$. Intuitively, the second type of data is inherently noisier as the margin is small and therefore, a classifier would struggle more to generalize on this type of data. We

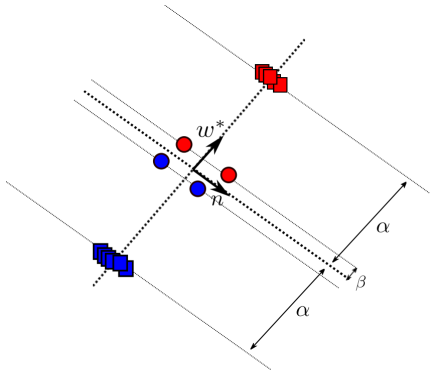


Figure 2: Dataset equation (D) 2D. Each data-point is $X_i = [c_i \cdot w^*, d_i \cdot n] \in \mathbb{R}^4$ for some $c_i, d_i \in \mathbb{R}$. We project these points in the 2D space $(\text{span}(w^*), \text{span}(n))$. The feature is w^* and the noisy patch is in $\text{span}(n)$. The large margin data (squares) have large component along w^* and relatively small noise component and are thus roughly equal to αw^*. The small margin data (circles) have relatively large noise component and thus, these data are well-spread on the span of n.

Student \ Teacher	Linear	1-MLP	2-MLP	1-CNN	2-CNN
1-MLP	93.48/93.25	92.32/92.18	84.3/83.68	94.18/94.12	76.04/76.12
2-MLP	93.45/92.85	91.02/91.78	83.82/83.25	94.14/94.20	75.50/75.56
1-CNN	92.21/92.34	92.31/92.33	83.39/83.44	94.39/94.39	79.44/78.32
2-CNN	91.04/91.22	91.51/91.56	82.44/82.12	93.91/93.79	80.86/78.56

Table 1: Test accuracy obtained using GD/GD+M on a Gaussian synthetic dataset trained using neural network with ReLU activations. The training dataset consists in 500 data points in dimension 30 and test set in 5000 points. The student networks are trained for 1000 epochs to ensure zero training error. The results are averaged over 3 runs and we only report the mean (see Appendix for complete table).

underline that all the examples share the *same feature* but differ in the intensity of the signal. We consider a training dataset of size N with the following split for $\hat{\mu} \ll 1$:

$$\begin{aligned} &(1-\hat{\mu})N \text{ datapoints are with large margin,} \\ &\hat{\mu}N \text{ datapoints are with small margin data.} \end{aligned} \tag{D}$$

Figure 2 sketches equation (D) in a 2D setting. We emphasize that datasets having similar features and different margins are common in the real-world. Examples include object-recognition datasets such as CIFAR (Krizhevsky et al., 2009) or Imagenet (Deng et al., 2009) (for example, the "wheel feature" of a car can be strong or weak depending on the orientation of the car). More specifically, we believe that the dataset (D) can be viewed as a simplified model of these object-recognition datasets. In this context, the following informal theorems characterize the generalization of the GD and GD+M models. They dramatically simplify Theorem 3.1 and Theorem 3.2 but highlight the intuitions behind our results.

Theorem 1.1 (Informal, GD+M)**.** *There exists a dataset of the form (D) with size N such that a two-layer (over-parameterized) convolutional network trained with GD+M:*

1. *initially only learns large margin data from the $(1-\hat{\mu})N$ examples.*
2. *has large historical gradients that contain the feature w^* present in small margin data.*
3. *keeps learning the feature in the small margin data using its momentum historical gradients.*

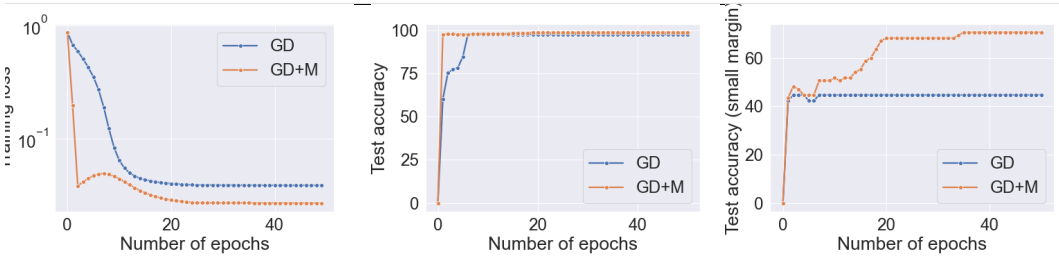
The model thus reaches zero training error and perfectly classify large and small margin data at test.

Theorem 1.2 (Informal, GD)**.** *There exists a dataset of the form (D) with size N such that a two-layer (over-parameterized) convolutional network trained with GD:*

1. *initially only learns large margin data from the from the $(1-\hat{\mu})N$ examples.*
2. *has small gradient after learning these data.*
3. *memorizes the remaining small margin data from the $\hat{\mu}N$ examples using the noises.*

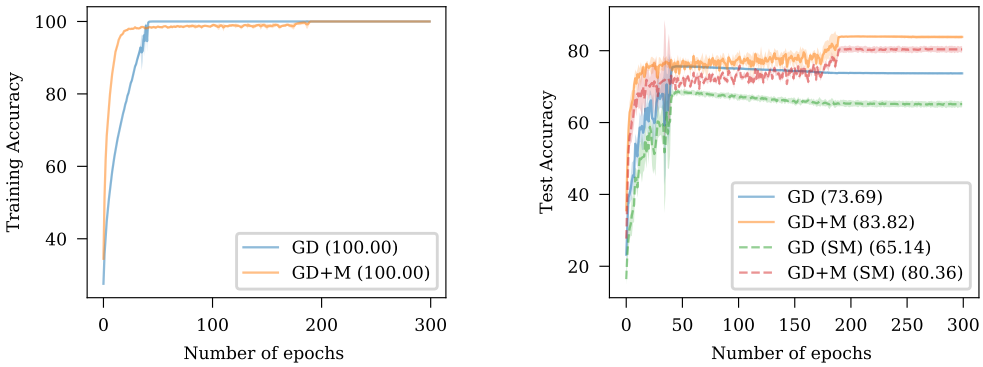
The model thus reaches zero training and manages to classify the large margin data at test. However, it fails to classify the small margin data because of the memorization step during training.

Why does GD+M generalize better than GD? Since the large margin data are dominant, GD focus in priority on these examples to decrease its training loss. However, after fitting this data, it significantly lowers its gradient. The gradient is thus not large enough for learning the small margin data. Similarly, GD+M fits the large margin data and subsequently gets a small gradient. However,



(a) (b) (c)

Figure 3: (a): Training loss (b) test accuracy on large margin data and (c) test accuracy on the small margin data in the synthetic setting. While GD and GD+M get zero training loss, GD+M generalizes better on small margin data than GD. Setting: 20000 training data, 2000 test data, d=30, number of neurons=5, number of patches=5.



(a) (b)

Figure 4: Training (a) and test (b) accuracy obtained with Resnet-18 on CIFAR-10 dataset with artificially generated small margin data. The architectures are trained using GD/GD+M for 300 epochs to ensure zero training error. Data augmentation, dropout and batch normalization are turned off. (SM) stands for the test accuracy obtained by the algorithm on the small margin data. Results are averaged over 5 runs with best scheduled learning rate and weight decay for each individual algorithm separately.

contrary to GD, GD+M has large historical gradients in his momentum gradient. These gradients helped to learn the feature in the large margin data. They also help to learn small margin data *since all the examples share the same feature*. GD+M therefore uses his momentum to learn the small margin data. We name this process *historical feature amplification* and believe that it is key to understand why momentum improves generalization.

Empirical justification. We also provide an empirical justification that such phenomenon does happen in a real-world setting as reported in Figure 4. In this experiment, we create small margin data in the CIFAR-10 dataset by respectively lowering the resolution of 10% of the training and test images, adding Gaussian noise of variance 0.005 and randomly shuffling the RGB channels. Figure 4 shows that even though both algortihms reach zero training error and 100% training accuracy, GD+M gets higher generalization than GD on this decimated dataset. Above all, at test, GD+M performs as well on small and large margin data while GD does relatively worse on small margin data.Indeed, the relative accuracy drop for GD+M is $80.36/83.32 = 0.97$ while for GD is equal to $65.14/73.69 = 0.88$.

Our paper is organized as follows. In Section 2, we formally define the data distribution equation (D), the model and algorithms we use to learn it. Lastly, Section 3 presents our main theorems and provide a proof sketch in Section 4 and Section 5. Additional experiments can be found in the Appendix.

More related Work

Momentum in convex setting. GD+M (a.k.a. heavy ball or Polyak momentum) consists in calculating the exponentially weighted average of the past gradients and using it to update the weights. For convex functions near a strict twice-differentiable minimum, GD+M is optimal regarding local convergence rate Polyak (1963; 1964); Nemirovskij & Yudin (1983); Nesterov (2003). However, it may fail to converge globally for general strongly convex twice-differentiable functions Lessard et al. (2015) and is no longer optimal for the class of smooth convex functions. In the stochastic setting, GD+M is more sensitive to noise in the gradients; that is, to preserve their improved convergence rates, significantly less noise is required d'Aspremont (2008); Schmidt et al. (2011); Devolder et al. (2014); Kidambi et al. (2018). Finally, other momentum methods are extensively used for convex functions such as Nesterov's accelerated gradient Nesterov (1983). Our paper focuses on the use of GD+M and contrary to the aforementioned papers, our setting is non-convex and we mainly focus on

the generalization of the model learned by GD and GD+M when both methods converge to global optimal. We underline that contrary to the non-convex world, generalization is typically disentangled with optimization for (strictly) convex functions.

Non-convex optimization with momentum. A long line of work consists in understanding the convergence speed of momentum methods when optimizing non-convex functions. Mai & Johansson (2020); Liu et al. (2020); Cutkosky & Mehta (2020); Defazio (2020) show that SGD+M reaches a stationary point as fast as SGD under diverse assumptions. Besides, Leclerc & Madry (2020) empirically shows that momentum accelerates neural network training for small learning rates and slows it down otherwise. Our paper differs from these works as we work in the batch setting and theoretically investigate the generalization benefits brought by momentum (and not the training ones).

Generalization with momentum. Momentum-based methods such as SGD+M, RMSProp (Tieleman & Hinton, 2012) and Adam (Kingma & Ba, 2014) are standard in deep learning training since the seminal work of Sutskever et al. (2013). Although its well accepted that Momentum improve generalization in deep learning, only a few works formally investigate the role of momentum in generalization. Leclerc & Madry (2020) *empirically* reports that momentum yields higher generalization when using a large learning rate. However, they assert that this benefit can be obtained by applying an even larger learning rate on vanilla SGD. We suspect that this observation is due to *batch normalization* (BN) which is known to dramatically bias the algorithm's generalization (Lyu & Li, 2019). In Appendix, we report that BN reduces the generalization gain of momentum comparing to without BN. To our knowledge, our work is first that *theoretically* investigate the generalization of momentum in deep learning.

2 Setting and Algorithms

In this section, we first introduce a formal definition of the data distribution equation (D) and the neural network model we use to learn it. We finally present the GD and GD+M algorithms.

General notations. For a matrix $W \in \mathbb{R}^{m\times d}$, we denote by w_r its r-th row. For a function $f\colon \mathbb{R}^{m\times d} \to \mathbb{R}$, we denote by $\nabla_{w_r} f(W)$ the gradient of f with respect to w_r and $\nabla f(W)$ the gradient with respect to W. For an optimization algorithm updating a vector w, $w^{(t)}$ represents its iterate at time t. We use $\mathbf{I}_d$ for the $d \times d$ identity matrix and $\mathbf{1}_m$ the all-ones vector of dimension m. Finally, we use the asymptotic complexity notations when defining the different constants in the paper. We use $\tilde{O}, \tilde{\Theta}, \tilde{\Omega}$ to hide logarithmic dependency on d.

Data distribution. We define our data distribution $\mathcal{D}$ as follows.

Each sample from $\mathcal{D}$ consists in an input data X and a label y that are generated as:

1. The label y is uniformly sampled from $\{-1, 1\}$.
2. Each data-point $X = (X[1], \dots, X[P])$ consists in P patches where each $X[j] \in \mathbb{R}^d$.
3. Signal patch: for one patch $P(X) \in [P]$, we have $X[P(X)] = cw^*$, where $c \in \mathbb{R}$, $w^* \in \mathbb{R}^d$ and $\|w^*\|_2 = 1$. ($\mathcal{D}$)
4. The distribution of c satisfies that
$$c = \begin{cases} \alpha y & \text{with probability } 1-\mu \\ \beta y & \text{with probability } \mu \end{cases}.$$
5. Noisy patches: for all the other patches $j \in [P]\backslash\{P(X)\}$, $X[j] \sim \mathcal{N}(0, (I - w^* w^{*\top})\sigma^2 \mathbf{I}_d)$.

We precise that we sample the noisy patches in the orthogonal complement of w^* to have a simpler analysis. To present the simplest result, we assume that the values in equation ($\mathcal{D}$) satisfy $\alpha = d^{0.49}$, $\beta = \frac{1}{\mathrm{polylog(d)}\sqrt{d}}\alpha$, $\sigma = \frac{1}{\sqrt{d}}$ and $P \in [2, \mathrm{polylog(d)}]$.

Using this model, we generate a training dataset $\mathcal{Z} = \{(X_i, y_i)\}_{i\in[N]}$ where $X_i = (X_i[j])_{j\in[P]}$. We focus on the case where $\mu = 1/\mathrm{poly}(d)$ and $N = \Theta\left(\frac{1}{\mu}\right)$. We let $\mathcal{Z}$ to be partitioned in two sets $\mathcal{Z}_1$ and $\mathcal{Z}_2$ such that $\mathcal{Z}_1$ gathers the large margin data while $\mathcal{Z}_2$ the small margin ones. Lastly, we define $\hat{\mu} = \frac{|\mathcal{Z}_2|}{N}$ the fraction of small margin data.

Learner model. We use a two-layer convolutional neural network with cubic activation to learn the training dataset $\mathcal{Z}$. This model is the simplest non-linear network since a quadratic activation

would only output positive labels and mismatch our labeling function. The first layer weights are $W \in \mathbb{R}^{m\times d}$ and the second layer is fixed to $\mathbf{1}_m$. Given a input data X, the output of the model is

$$f_W(X) = \sum_{r=1}^{m}\sum_{j=1}^{P}\langle w_r, X[j]\rangle^3. \tag{CNN}$$

The number of neurons is set as $m = \text{polylog}(d)$ to ensure that (CNN) is mildly over-parametrized.

Training objective. We fit the training dataset $\mathcal{Z}$ using (CNN) and solve the logistic regression problem

$$\min_{W\in\mathbb{R}^{m\times d}} \frac{1}{N}\sum_{i=1}^{N}\log\left(1+\exp\left(-y_i f_W(X_i)\right)\right) + \frac{\lambda}{2}\|W\|_2^2 := \widehat{L}(W). \tag{P}$$

(P) sheds light on our choice of cubic activation in (CNN). Indeed, it is the smallest polynomial degree that makes the training objective (P) non-convex and compatible with our dataset. Linear or quadratic activations would respectively make the problem convex or all the labels positive. Here, we pick $\lambda \in \left[0, \frac{1}{\text{poly}(d)N}\right]$.

Importance of non-convexity. When $\lambda > 0$, if the loss $\frac{1}{N}\sum_{i=1}^{N}\log\left(1+\exp\left(-y_i f_W(X_i)\right)\right)$ is convex, then there is a unique global optimal solution, so the choice of optimization algorithm **does not matter**. In our case, due to the non-convexity of the training objective, GD + M converges to a different (approximate) global optimal comparing to GD, with better generalization properties.

Test error. We assess the quality of a predictor $\widehat{W}$ using the classical 0-1 loss used in binary classification. Given a sample (X, y), the *individual test (classification) error* is defined as $\mathscr{L}(X, y) = \mathbf{1}\{f_{\widehat{W}}(X)y < 0\}$. While $\mathscr{L}$ measures the error of f_W on an individual data-point, we are interested in the *test error* that measures the average loss over data points generated from ($\mathcal{D}$) and defined as

$$\mathscr{L}(f_{\widehat{W}}) := \mathbb{E}_{(X,y)\sim\mathcal{D}}[\mathscr{L}(f_{\widehat{W}}(X), y)]. \tag{TE}$$

Algorithms. We solve the training problem equation (P) using GD and GD+M. GD is defined by

$$W^{(t+1)} = W^{(t)} - \eta\nabla\widehat{L}(W^{(t)}), \quad \textit{for } t \geq 0, \tag{GD}$$

where $\eta > 0$ is the learning rate. On the other hand, GD+M is defined by the update rule

$$\begin{cases} g^{(t+1)} &= \gamma g^{(t)} + (1-\gamma)\nabla\widehat{L}(W^{(t)}) \\ W^{(t+1)} &= W^{(t)} - \eta g^{(t+1)} \end{cases}, \quad \textit{for } t \geq 0. \tag{GD+M}$$

where $\gamma \in (0, 1)$ is momentum factor. We now detail how to set parameters in (GD) and (GD+M).

Parametrization 2.1. *When running GD and GD+M on equation (P), the number of iterations is* $T \in \left[\text{poly(d)}N/(\eta), d^{O(\log d)}/(\eta)\right]$. *For both algorithms, the weights* $w_1^{(0)}, \ldots, w_m^{(0)}$ *are initialized using independent samples from a normal distribution* $\mathcal{N}(0, \sigma_0^2\mathbf{I}_d)$ *where* $\sigma_0^2 = \frac{\text{polylog}(d)}{d}$. *The learning rate is set as:*

1. *GD: the learning rate may take any reasonable value* $\eta \in (0, \tilde{O}(1)]$.
2. *GD+M: the learning rate is a large learning rate:* $\eta = \tilde{\Theta}(1)$.[1]

Lastly, the momentum factor in GD+M is set to be $\gamma = 1 - \frac{\text{polylog}(d)}{d}$.

Our Parametrization 2.1 matches with the parameters used in practice as the weights are generally initialized from Gaussian with small variance and momentum is set close to 1 (Sutskever et al., 2013).

3 Main Results

We now formally state our main theorems regarding the generalization of models trained using equation (GD) and equation (GD+M) on the training set $\mathcal{Z}$ generated by equation ($\mathcal{D}$). As announced in the introduction, we show that the GD+M model incurs a generalization error that is dramatically smaller than the GD model. Before introducing the main result, we define some notations:

[1] This is consistent with the empirical observation that only momentum with large learning rate improves generalization (Sutskever et al., 2013)

Main objects. Let $r \in [m]$, $i \in [N]$, $j \in P \backslash \{P(X_i)\}$, $\gamma \in (0,1)$ and $t \geq 0$. We are mainly interested in $w_r^{(t)}$, the r-th weight of the network, $\nabla_{w_r} \widehat{L}(W^{(t)})$ the gradient of the training loss w.r.t. w_r, $g_r^{(t)}$ the momentum gradient defined by $g_r^{(t+1)} = \gamma g_r^{(t)} + (1-\gamma)\nabla_{w_r} \widehat{L}(W^{(t)})$. The analysis lies on the projection of these objects on the feature w^* and on noisy patches $X_i[j]$. We introduce the following notations for the component of the learned weights along feature and noise directions:

– Projection on w^*: $c_r^{(t)} = \langle w_r^{(t)}, w^* \rangle$.

– Projection on $X_i[j]$: $\Xi_{i,j,r}^{(t)} = \langle w_r^{(t)}, X_i[j] \rangle$.

– Total noise: $\Xi_i^{(t)} = \sum_{r=1}^m \sum_{j \in [P] \backslash \{P(X_i)\}} \langle w_r^{(t)}, X_i[j] \rangle^3$.

– Maximum signal: let $r_{\max} = \operatorname{argmax}_{r \in [m]} c_r^{(t)}$, $c^{(t)} = c_{r_{\max}}^{(t)}$

Theorem 3.1. *Assume that we run GD on (P) for T iterations with parameters set as in Parametrization 2.1. With probability at least $1 - o(1)$, the weights learned by GD*

1. *partially learn the feature: for all $r \in [m]$, $|c_r^{(T)}| \leq \tilde{O}(1/\alpha)$.*
2. *memorize from small margin data: for all $i \in \mathcal{Z}_2$, $\Xi_i^{(t)} \geq \tilde{\Omega}(1)$.*

*Consequently, the training error is smaller than $O(\mu/\mathrm{poly}(d))$ and the test error is **at least** $\tilde{\Omega}(\mu)$.*

Intuitively, the training process of the GD model is described as follows. Since the large margin data are dominant in $\mathcal{Z}$, the gradient points mainly in the direction of the feature w^*. Therefore, GD eventually learns the feature in $\mathcal{Z}_1$ (Lemma 4.1) and the gradients from $\mathcal{Z}_1$ quickly become small. Afterwards, the gradient is dominated by the gradients from $\mathcal{Z}_2$ (Lemma 4.2). Because $\mathcal{Z}_2$ has small margin, the full gradient is now directed by the noisy patches. It implies that GD memorizes noise in $\mathcal{Z}_2$ (Lemma 4.4). Since these gradients also control the amount of remaining feature to be learned (Lemma 4.3), we conclude that the GD model partially learns the feature and introduces a huge noise component in the learned weights. We provide a proof sketch of Theorem 3.1 in Section 4.

Theorem 3.2. *Assume that we run GD+M on equation (P) for T iterations with parameters set as in Parametrization 2.1. With probability at least $1 - o(1)$, the weights learned by GD+M*

1. *(at least for one of them) is highly correlated with the feature: $c^{(T)} > \tilde{\Omega}(1/\beta)$.*
2. *are barely correlated with noise: for all $r \in [m]$, for all $i \in [N]$ and $j \in [P]$. $|\Xi_{i,j,r}^{(T)}| \leq \tilde{O}(\sigma_0)$.*

*Consequently, the training loss and the test error are **at most** $O(\mu/\mathrm{poly}(d))$.*

Intuitively, the GD+M model follows this training process. Similarly to GD, it first fits the $\mathcal{Z}_1$ (Lemma 5.1). Contrary to GD, the momentum gradient is still highly correlated with w^* after this step (Lemma 5.2). Indeed, the key difference is that momentum accumulates historical gradients. Since these gradients were accumulated when learning large margin data, the direction of momentum gradient is highly biased towards w^*. Therefore, the GD+M model *amplifies the feature* from these historical gradients to learn the feature in small margin data (Lemma 5.3). Subsequently, the gradient becomes small (Lemma 5.4) and the weights are no longer updated. Therefore, the GD+M model manages to ignore the noisy patches (Lemma 5.5) and learns the feature from both $\mathcal{Z}_1$ and $\mathcal{Z}_2$. We provide a proof sketch of Theorem 3.2 in Section 5.

To state the proof, we further decompose the gradients along signal and noise directions.

– Projection on w^*: $\mathscr{G}_r^{(t)} = \langle \nabla_{w_r} \widehat{L}(W_t), w^* \rangle$ and $\mathcal{G}_r^{(t)} = \langle g_r^{(t)}, w^* \rangle$.

– Projection on $X_i[j]$: $\mathsf{G}_{i,j,r}^{(t)} = \langle \nabla_{w_r} \widehat{L}(W^{(t)}), X_i[j] \rangle$, $G_{i,j,r}^{(t)} = \langle g_r^{(t)}, X_i[j] \rangle$.

– Maximum signal: let $r_{\max} = \operatorname{argmax}_{r \in [m]} c_r^{(t)}$, $c^{(t)} = c_{r_{\max}}^{(t)}$ and $\mathcal{G}^{(t)} = \mathcal{G}_{r_{\max}}^{(t)}$.

Signal and noise iterates. Our analysis is build upon a decomposition of the updates equation (GD) and equation (GD+M) on w^* and $X_i[j]$. These decompositions are respectively defined as follows:

$$c_r^{(t+1)} = c_r^{(t)} - \eta \mathscr{G}_r^{(t)} \quad \text{(GD-S)} \qquad \Xi_{i,j,r}^{(t+1)} = \Xi_{i,j,r}^{(t)} - \eta \mathsf{G}_{i,j,r}^{(t)} \quad \text{(GD-N)}$$

$$\begin{cases} \mathcal{G}_r^{(t+1)} = \gamma \mathcal{G}_r^{(t)} + (1-\gamma)\mathscr{G}_r^{(t)} \\ c_r^{(t+1)} = c_r^{(t)} - \eta \mathcal{G}_r^{(t+1)} \end{cases} \text{(GDM-S)} \qquad \begin{cases} G_{i,j,r}^{(t+1)} = \gamma G_{i,j,r}^{(t)} + (1-\gamma)\mathbb{G}_{i,j,r}^{(t)} \\ \Xi_{i,j,r}^{(t+1)} = \Xi_{i,j,r}^{(t)} - G_{i,j,r}^{(t+1)} \end{cases} \text{(GDM-N)}$$

We detail how to use these dynamics to analyze GD+M and GD in Section 4 and Section 5. Our analysis heavily depends on the gradients of the training loss which involve $\text{sigmoid}(x) = (1+e^{-x})^{-1}$. We define the derivative of a data-point i as $\ell_i^{(t)} = \text{sigmoid}(-y_i f_{W^{(t)}}(X_i))$, the derivatives $\nu_k^{(t)} = \frac{1}{N}\sum_{i\in\mathcal{Z}_k} \ell_i^{(t)}$ for $k \in \{1,2\}$ and the full derivative $\nu^{(t)} = \nu_1^{(t)} + \nu_2^{(t)}$.

4 Analysis of GD

In this section, we provide a proof sketch for Theorem 3.1 that reflects the behavior of GD with $\lambda = 0$. A more detailed proof (extending to $\lambda > 0$) can be found in the Appendix.

Step 1: Learning $\mathcal{Z}_1$. At the beginning of the learning process, the gradient is mostly dominated by the gradients coming from the $\mathcal{Z}_1$ samples. Since these data have large margin, the gradient is thus highly correlated with w^* and $c_r^{(t)}$ increases as shown in the following Lemma.

Lemma 4.1. *For all $r \in [m]$ and $t \geq 0$, equation (GD-S) is simplified as:*

$$c_r^{(t+1)} \geq c_r^{(t)} + \tilde{\Theta}(\eta)\alpha^3 (c_r^{(t)})^2 \cdot \text{sigmoid}(-\textstyle\sum_{s=1}^t \alpha^3 (c_s^{(t)})^3).$$

Consequently, after $T_0 = \tilde{\Theta}\left(\frac{1}{\eta\alpha^3\sigma_0}\right)$ iterations, for all $t \in [T_0, T]$, we have $c^{(t)} \geq \tilde{\Omega}(1/\alpha)$.

Intuitively, the increment in the update in Lemma 4.1 is non-zero when the sigmoid is not too small which is equivalent to $c^{(t)} \leq \tilde{O}(1/\alpha)$. Therefore, $c^{(t)}$ keeps increasing until reaching this threshold. After this step, the $\mathcal{Z}_1$ data have small gradient and therefore, GD has learned these data.

Lemma 4.2. *Let $T_0 = \tilde{\Theta}\left(\frac{1}{\eta\alpha^3\sigma_0}\right)$. After $t \in [T_0, T]$ iterations, the $\mathcal{Z}_1$ derivative is bounded as $\nu_1^{(t)} \leq \tilde{O}\left(\frac{1}{\eta(t-T_0+1)\alpha}\right) + \tilde{O}\left(\frac{\beta^3}{\alpha}\right)\nu_2^{(t)}$. The full derivative is $\nu^{(t)} \leq \tilde{O}\left(\frac{1}{\eta(t-T_0+1)\alpha} + \left(1+\frac{\beta^3}{\alpha}\right)\nu_2^{(t)}\right)$.*

By our choice of parameter, Lemma 4.2 indicates that the full gradient is dominated by the gradients from $\mathcal{Z}_2$ data after $T_0 = \tilde{\Omega}\left(\frac{1}{\tilde{\mu}\eta\alpha}\right)$. Consequently, $\nu_2^{(t)}$ also rules the amount of feature learnt by GD.

Lemma 4.3. *Let $T_0 = \tilde{\Theta}\left(\frac{1}{\eta\alpha^3\sigma_0}\right)$. For $t \in [T_0, T]$, equation (GD-S) becomes $c^{(t+1)} \leq \tilde{O}(1/\alpha) + \tilde{O}(\eta\beta^3/\alpha)\sum_{\tau=T_0}^t \nu_2^{(\tau)}$.*

Lemma 4.3 implies that quantifying the decrease rate of $\nu_2^{(t)}$ provides an estimate on the quantity of feature learnt by the model. We remark that $\nu_2^{(t)} = \text{sigmoid}(\beta^3 \sum_{s=1}^m (c_s^{(t)})^3 + \Xi_i^{(t)})$ for some $i \in \mathcal{Z}_2$. We thus need to determine whether the feature or the noise terms dominates in the sigmoid.

Step 2: Memorizing $\mathcal{Z}_2$. We now show that the total correlation between the weights and the noise in $\mathcal{Z}_2$ data increases until being large.

Lemma 4.4. *Let $t \geq 0$ and $i \in \mathcal{Z}_2$. Assume that $\Xi_i^{(t)} \leq \tilde{O}(1)$. Then, equation (GD-N) can be simplified as:*

$$y_i \Xi_{i,j,r}^{(t+1)} \geq y_i \Xi_{i,j,r}^{(0)} + \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N}\sum_{\tau=0}^t (\Xi_{i,j,r}^{(\tau)})^2 - \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right).$$

Let $T_1 = \tilde{O}\left(\frac{N}{\sigma_0\sigma\sqrt{d}\sigma^2 d}\right)$. Therefore, $\Xi_i^{(t)} \geq \tilde{\Omega}(1)$, for $t \in [T_1, T]$ and thus GD memorizes.

By Lemma 4.4, in the gradient of $\mathcal{Z}_2$ data, the noise term dominates the feature term (which scales as $\tilde{O}(\beta^3)$). Consequently, the algorithm memorizes the $\mathcal{Z}_2$ data which implies a fast decay of $\nu_2^{(t)}$.

Lemma 4.5. *Let* $T_1 = \tilde{O}\left(\frac{N}{\sigma_0 \sigma \sqrt{d \sigma^2 d}}\right)$. *For* $t \in [T_1, T]$, *we have* $\sum_{\tau=0}^{t} \nu_2^{(\tau)} \leq \tilde{O}\left(\frac{1}{\eta \sigma_0}\right)$.

Combining Lemma 4.5 and Lemma 4.3, we prove that GD partially learns the feature.

Lemma 4.6. *For* $t \leq T$, *the signal component satisfies* $c^{(t)} \leq \tilde{O}(1/\alpha)$.

Lemma 4.4 and Lemma 4.6 respectively yield the first two items in Theorem 3.1. Bounds on the training and test errors are respectively obtained by plugging these results in (P) and (TE).

5 Analysis of GD+M

In this section, we provide a proof sketch for Theorem 3.2 that reflects the behavior of GD+M with $\lambda = 0$. A more detailed proof (also extending to $\lambda > 0$) can be found in the Appendix.

Step 1: Learning $\mathcal{Z}_1$. Similarly to GD, by our initialization choice, the early gradients and so, momentum gradients are large. They are also spanned by the feature w^* and therefore, the GD+M model also increases its correlation with w^*.

Lemma 5.1. *For all* $r \in [m]$ *and* $t \geq 0$, *as long as* $c^{(t)} \leq \tilde{O}(1/\alpha)$, *the momentum update equation (GDM-S) is simplified as:*

$$-\mathcal{G}_r^{(t+1)} = -\gamma \mathcal{G}_r^{(t)} + (1-\gamma)\Theta(\alpha^3)(c_r^{(t)})^2$$

Consequently, after $\mathcal{T}_0 = \tilde{\Theta}\left(\frac{1}{\sigma_0 \alpha^2} + \frac{1}{1-\gamma}\right)$ *iterations, for all* $t \in [\mathcal{T}_0, T]$, *we have* $c^{(t)} \geq \tilde{\Omega}(1/\alpha)$.

Step 2: Learning $\mathcal{Z}_2$. Contrary to GD, GD+M has a large momentum that contains w^* after Step 1.

Lemma 5.2. *Let* $\mathcal{T}_0 = \tilde{\Theta}\left(\frac{1}{\sigma_0 \alpha^3} + \frac{1}{1-\gamma}\right)$. *For* $t \in [\mathcal{T}_0, T]$, *we have* $\mathcal{G}^{(t)} \geq \tilde{\Omega}(\sqrt{1-\gamma}/\alpha)$.

Lemma 5.2 hints an important distinction between GD and GD+M: while the current gradient along w^* is small at time $\mathcal{T}_0$, the momentum gradient stores historical gradients that are spanned by w^*. It *amplifies* the feature present in previous gradients to learn the feature from small margin data.

Lemma 5.3. *Let* $\mathcal{T}_0 = \tilde{\Theta}\left(\frac{1}{\sigma_0 \alpha^3} + \frac{1}{1-\gamma}\right)$. *After* $\mathcal{T}_1 = \mathcal{T}_0 + \tilde{\Theta}\left(\frac{1}{1-\gamma}\right)$ *iterations, for* $t \in [\mathcal{T}_1, T]$, *we have* $c^{(t)} \geq \tilde{\Omega}\left(\frac{1}{\sqrt{1-\gamma}\alpha}\right)$. *Our choice of parameter in Section 2, this implies* $c^{(t)} \geq \tilde{\Omega}(1/\beta)$.

Lemma 5.3 states that at least one of the weights that is highly correlated with the feature compared to GD where $c^{(t)} = \tilde{O}(1)$. This result implies that $\nu^{(t)}$ converges fast.

Lemma 5.4. *Let* $\mathcal{T}_0 = \tilde{\Theta}\left(\frac{1}{\eta \sigma_0 \alpha^3} + \frac{1}{1-\gamma}\right)$. *After* $\mathcal{T}_1 = \mathcal{T}_0 + \tilde{\Theta}\left(\frac{1}{1-\gamma}\right)$ *iterations, for* $t \in [\mathcal{T}_1, T]$, $\nu^{(t)} \leq \tilde{O}\left(\frac{1}{\eta(t-\mathcal{T}_1+1)\beta}\right)$.

With this fast convergence, Lemma 5.4 implies that the correlation of the weights with the noisy patches does not have enough time to increase and thus, remains small.

Lemma 5.5. *Let* $i \in [N]$, $j \in [P] \backslash \{P(X_i)\}$ *and* $r \in [m]$. *For* $t \geq 0$, *equation (GDM-N) can be rewritten as* $|G_{i,j,r}^{(t+1)}| \leq \gamma |G_{i,j,r}^{(t)}| + (1-\gamma)\tilde{O}(\sigma_0^2 \sigma^4 d^2)\nu^{(t)}$. *As a consequence, after* $t \in [\mathcal{T}_1, T]$ *iterations, we thus have* $|\Xi_{i,j,r}^{(t)}| \leq \tilde{O}(\sigma_0 \sigma \sqrt{d})$.

Lemma 5.3 and Lemma 5.5 respectively yield the two first items in Theorem 3.2.

6 Discussion

Our work is a first step towards understanding the algorithmic regularization of momentum and leaves room for improvements. We constructed a data distribution where historical feature amplification may explain the generalization improvement of momentum. However, it would be interesting to understand whether this phenomenon is the only reason or whether there are other mechanisms explaining momentum's benefits.An interesting setting for this question is NLP where momentum is used to train large models as BERT (Devlin et al., 2018). Lastly, our analysis is in the batch setting to isolate the generalization induced by momentum. It would be interesting to understand how the stochastic noise and the momentum together contribute to the generalization of a neural network.

REFERENCES

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

Sanjeev Arora, Zhiyuan Li, and Kaifeng Lyu. Theoretical analysis of auto rate-tuning by batch normalization. *arXiv preprint arXiv:1812.03981*, 2018.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *arXiv preprint arXiv:1905.13655*, 2019.

Anthony Carbery and James Wright. Distributional and l^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical research letters*, 8(3):233–248, 2001.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pp. 1305–1338. PMLR, 2020.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized sgd. In *International Conference on Machine Learning*, pp. 2260–2268. PMLR, 2020.

Alexandre d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008.

Aaron Defazio. Understanding the role of momentum in non-convex optimization: Practical insights from a lyapunov analysis. *arXiv preprint arXiv:2010.00406*, 2020.

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

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

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75, 2014.

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

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–10. IEEE, 2018.

Jeff Z HaoChen, Colin Wei, Jason D Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. *arXiv preprint arXiv:2006.08680*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *arXiv preprint arXiv:1705.08741*, 2017.

Elad Hoffer, Ron Banner, Itay Golan, and Daniel Soudry. Norm matters: efficient and accurate normalization schemes in deep networks, 2019.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pp. 1772–1798. PMLR, 2019.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9. IEEE, 2018.

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

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Guillaume Leclerc and Aleksander Madry. The two regimes of deep network training. *arXiv preprint arXiv:2002.10376*, 2020.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints, 2015.

Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *arXiv preprint arXiv:1907.04595*, 2019.

Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. *arXiv preprint arXiv:2007.07989*, 2020.

Shachar Lovett. An elementary proof of anti-concentration of polynomials in gaussian variables. *Electron. Colloquium Comput. Complex.*, 17:182, 2010.

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

Vien Mai and Mikael Johansson. Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *International Conference on Machine Learning*, pp. 6630–6639. PMLR, 2020.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady an ussr*, volume 269, pp. 543–547, 1983.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-sgd: Path-normalized optimization in deep neural networks. *arXiv preprint arXiv:1506.02617*, 2015.

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. *arXiv preprint arXiv:1109.2415*, 2011.

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

Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. Don't decay the learning rate, increase the batch size, 2018.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1): 2822–2878, 2018.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147. PMLR, 2013.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout, 2020.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *arXiv preprint arXiv:1705.08292*, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

A ADDITIONAL EXPERIMENTS

In this section, we present more details on the experiments presented in the introduction and additional experiments.

A.1 GD/GD+M AND SGD/SGD+M ON CIFAR-10 AND CIFAR-100

We start by giving a complete table with the mean and standard deviations obtained by running a Resnet18 and a WideResnet16 on CIFAR-10 and CIFAR-100. This table completes the one given in Section 1.

	CIFAR-10	CIFAR-100
Resnet18	76.27 ± 0.15	43.32 ± 0.13
WideResnet16	75.02 ± 0.20	42.95 ± 0.21

(a)

	CIFAR-10	CIFAR-100
Resnet18	85.20 ± 0.10	51.99 ± 0.15
WideResnet16	84.48 ± 0.28	51.33 ± 0.23

(b)

	CIFAR-10	CIFAR-100
Resnet18	86.15 ± 0.11	53.81 ± 0.32
WideResnet16	84.83 ± 0.21	55.09 ± 0.45

(c)

	CIFAR-10	CIFAR-100
Resnet18	85.91 ± 0.21	58.01 ± 0.28
WideResnet16	87.85 ± 0.19	60.83 ± 0.25

(d)

Table 2: Test accuracy obtained with Resnet-18 and WideResnet16 on CIFAR-10 and CIFAR-100. The architectures are trained using GD (a) GD+M (b), SGD (c) and SGD+M (d) for 300 epochs. To isolate the effect of momentum, we *turn off* data augmentation, dropout and batch normalization. GD and SGD respectively refer to stochastic gradient descent with batch sizes 1024 and 128. The results are averaged over 3 runs.

A.2 INFLUENCE OF THE BATCH-SIZE ON THE GENERALIZATION IMPROVEMENT OF MOMENTUM

In this section, we study the relationship between the size of the batch-size and the generalization improvement induced by momentum. Table 3 confirms on larger spectrum of batch-sizes the observation made in the introduction: momentum induces a more significant improvement in terms of test accuracy for large batch-sizes.

	16	128	1024	2048
momentum	87.26 ± 0.35	85.91 ± 0.21	85.20 ± 0.10	84.73 ± 0.11
no momentum	86.62 ± 0.28	86.15 ± 0.11	76.27 ± 0.15	74.96 ± 0.08

(a)

	16	128	1024	2048
momentum	62.10 ± 0.35	58.01 ± 0.28	53.81 ± 0.32	52.96 ± 0.13
no momentum	56.25 ± 0.28	51.99 ± 0.15	43.32 ± 0.13	41.73 ± 0.15

(b)

Table 3: Test accuracy obtained with Resnet-18 on CIFAR-10 (a) and CIFAR-100 (b). The architectures are trained using GD with varying batch-sizes (16,128,1024,2048) for 300 epochs. To isolate the effect of momentum, we *turn off* data augmentation, dropout and batch normalization. The results are averaged over 3 runs.

A.3 SYNTHETIC GAUSSIAN DATA EXPERIMENTS

We provide a complete table with mean and standard deviations obtained by using different student networks to learn a synthetic dataset. This datasets consists of Gaussian data-points and the labels are generated by teacher networks with varying complexity. Table 4 and Table 5 show that momentum does not help getting a higher generalization in the batch or stochastic settings.

Student \ Teacher	Linear	1-MLP	2-MLP	1-CNN	2-CNN
1-MLP	93.48 ± 0.13	92.32 ± 0.50	84.30 ± 0.82	94.18 ± 0.42	76.04 ± 0.29
2-MLP	93.45 ± 0.22	91.02 ± 0.41	83.82 ± 0.43	94.14 ± 0.47	75.50 ± 0.35
1-CNN	92.21 ± 0.16	92.31 ± 0.57	83.39 ± 0.48	94.39 ± 0.17	79.44 ± 0.58
2-CNN	91.04 ± 0.48	91.51 ± 0.40	82.44 ± 0.45	93.91 ± 0.35	80.86 ± 0.92

(a)

Student \ Teacher	Linear	1-MLP	2-MLP	1-CNN	2-CNN
1-MLP	93.25 ± 0.22	92.18 ± 0.53	83.68 ± 0.74	94.12 ± 0.43	76.12 ± 0.22
2-MLP	92.85 ± 0.34	91.78 ± 0.62	83.25 ± 0.70	94.20 ± 0.13	75.56 ± 0.33
1-CNN	92.34 ± 0.21	92.33 ± 0.64	83.44 ± 0.52	94.39 ± 0.15	78.32 ± 0.34
2-CNN	91.22 ± 0.39	91.56 ± 0.52	82.12 ± 0.55	93.79 ± 0.25	78.56 ± 0.64

(b)

Table 4: Test accuracy obtained using GD (a) and GD+M (b) on a Gaussian synthetic dataset trained using neural network with ReLU activations. The training dataset consists in 500 data points in dimension 50 and test set in 5000 points. The student networks are trained for 1000 epochs to ensure zero training error. The results are averaged over 3 runs.

Student \ Teacher	Linear	1-MLP	2-MLP	1-CNN	2-CNN
1-MLP	93.58 ± 0.32	92.56 ± 0.62	85.74 ± 0.56	94.18 ± 0.42	76.06 ± 0.39
2-MLP	93.51 ± 0.25	91.82 ± 0.83	85.33 ± 0.81	94.14 ± 0.33	75.33 ± 0.47
1-CNN	92.42 ± 0.05	92.03 ± 0.53	84.57 ± 0.47	94.22 ± 0.18	80.02 ± 0.45
2-CNN	91.54 ± 0.37	92.04 ± 0.48	83.81 ± 0.47	93.95 ± 0.31	82.86 ± 0.59

(c)

Student \ Teacher	Linear	1-MLP	2-MLP	1-CNN	2-CNN
1-MLP	93.56 ± 0.28	92.82 ± 0.26	84.65 ± 0.45	94.16 ± 0.42	76.01 ± 0.33
2-MLP	93.24 ± 0.34	92.26 ± 0.76	84.27 ± 0.79	94.24 ± 0.40	75.04 ± 0.47
1-CNN	92.50 ± 0.05	91.68 ± 0.72	83.39 ± 0.44	94.07 ± 0.035	78.92 ± 0.41
2-CNN	91.61 ± 0.41	91.94 ± 0.54	83.70 ± 0.37	93.89 ± 0.33	80.50 ± 0.45

(d)

Table 5: Test accuracy obtained using SGD (c) and GD+M (d) on a Gaussian synthetic dataset trained using neural network with ReLU activations. The training dataset consists in 500 data points in dimension 50 and test set in 5000 points. The student networks are trained for 1000 epochs to ensure zero training error. The results are averaged over 3 runs.

A.4 Momentum and batch normalization

In this section, we present the test error achieved by a VGG-16 and Resnet-18 trained with batch normalization on CIFAR-10 and CIFAR-100. We precise that contrary to the introduction, we do not train a WideResnet16 because of our limited memory. Table 6 indicates that the batch normalization drastically reduces the generalization improvement induced by momentum. We observe a slight improvement for large batch sizes but no significant improvement for small ones. These tables match with the observations made in Leclerc & Madry (2020) who assert that momentum does not improve generalization when the architecture is trained with batch normalization.

	CIFAR-10	CIFAR-100
VGG16	83.13 ± 0.10	53.81 ± 0.15
Resnet18	85.20 ± 0.28	55.75 ± 0.23

(a)

	CIFAR-10	CIFAR-100
VGG16	87.33 ± 0.11	60.28 ± 0.19
Resnet18	88.78 ± 0.07	61.32 ± 0.09

(b)

	CIFAR-10	CIFAR-100
VGG16	88.15 ± 0.08	59.83 ± 0.12
Resnet18	89.21 ± 0.21	62.47 ± 0.19

(c)

	CIFAR-10	CIFAR-100
VGG16	89.64 ± 0.05	58.01 ± 0.10
Resnet18	90.10 ± 0.08	63.17 ± 0.12

(d)

Table 6: Test accuracy obtained with VGG-16 and Resnet-18 with batch-normalization on CIFAR-10 and CIFAR-100. The architectures are trained using GD (a) GD+M (b), SGD (c) and SGD+M (d) for 300 epochs. To isolate the effect of momentum, we *turn off* data augmentation, dropout. GD and SGD respectively refer to stochastic gradient descent with batch sizes 1024 and 128. The results are averaged over 3 runs.

B NOTATIONS

In this section, we introduce the different notations used in the proofs. We start by defining the notations that appear for GD and GD+M. We first consider the case when $\lambda = 0$, we will extend the proof to $\lambda > 0$ in section (G)

B.1 NOTATIONS FOR GD AND GD+M

Our paper rely on the notions of signal and noise components of the iterates.

– Signal intensity: $\theta = \alpha$ if $i \in \mathcal{Z}_1$ and β otherwise.

– Signal: $c_r^{(t)} = \langle w^*, w_r^{(t)} \rangle$ for $r \in [m]$.

– Max signal: $c^{(t)} = c_{r_{\max}}^{(t)}$ where $r_{\max} \in \operatorname{argmax}_{r \in [m]} c_r^{(t)}$.

– Noise: $\Xi_{i,j,r}^{(t)} = \langle w_r^{(t)}, X_i[j] \rangle$ for $i \in [N]$ and $j \in [P] \backslash \{P(X_i)\}$.

– Max noise: $\Xi_{\max}^{(t)} = \max_{i \in [N], j \neq P(X^{(i)}), r \in [m]} |\Xi_{i,j,r}^{(t)}|^2$.

– Total noise: $\Xi_i^{(t)} = y_i \sum_{r \in [m], j \in [P], j \neq P(X^{(i)})} \left(\Xi_{i,j,r}^{(t)} \right)^3$.

We also use the following notations when dealing with the loss function and its gradient.

– Signal loss: $\widehat{\mathcal{L}}^{(t)}(a) = \log \left(1 + \exp \left(- \sum_{r=1}^m (c_r^{(t)})^3 a^3 \right) \right)$ for $a \in \mathbb{R}$.

– Noise loss: $\widehat{\mathcal{L}}^{(t)}(\Xi_i^{(t)}) = \log \left(1 + \exp \left(-\Xi_i^{(t)} \right) \right)$.

– Negative sigmoid function: $\mathfrak{S}(x) = (1 + \exp(x))^{-1}$, for $x \in \mathbb{R}$.

– Signal derivative: $\widehat{\ell}^{(t)}(a) = \mathfrak{S} \left(- \sum_{r=1}^m (c_r^{(t)})^3 a^3 \right)$, for $a \in \mathbb{R}$.

– Noise derivative: $\widehat{\ell}^{(t)}(\Xi_i^{(t)}) = \mathfrak{S}(-\Xi_i^{(t)})$.

– Derivative: $\ell_i^{(t)} = \mathfrak{S} \left(- \sum_{r=1}^m \sum_{j=1}^P \langle w_r^{(t)}, X_i[j] \rangle^3 \right)$, for $i \in [N]$.

– $\mathcal{Z}_k$ derivative: $\nu_k^{(t)} = \frac{1}{N} \sum_{i \in \mathcal{Z}_k} \ell_i^{(t)}$ for $k \in \{1, 2\}$.

– Full derivative: $\nu^{(t)} = \nu_1^{(t)} + \nu_2^{(t)}$.

– Gradient on signal: $\mathscr{G}_r^{(t)} = \langle \nabla_{w_r} \widehat{L}(W^{(t)}), w^* \rangle$ for $r \in [m]$.

– Gradient on noise: $\mathbb{G}_{i,j,r}^{(t)} = \langle \nabla_{w_r} \widehat{L}(W^{(t)}), X_i[j] \rangle$ for $i \in [N]$, $j \in [P] \backslash \{P(X_i)\}$ and $r \in [m]$.

– Gradient on normalized noise: $\mathrm{G}_r^{(t)} = \left\langle \nabla_{w_r} \widehat{L}(W^{(t)}), \chi \right\rangle$, for $r \in [m]$, where $\chi = \frac{\frac{1}{N} \sum_{i \in \mathcal{Z}_2} \sum_{j \neq P(X_i)} X_i[j]}{\| \frac{1}{N} \sum_{i \in \mathcal{Z}_2} \sum_{j \neq P(X_i)} X_i[j] \|_2}$.

B.2 Notations specific to GD+M

We now introduce the notations that only appear in the proofs involving GD+M.

– Momentum gradient oracle: $g_r^{(t)} = \gamma g_r^{(t-1)} + (1-\gamma) \nabla_{w_r} \widehat{L}(W^{(t)})$ for $r \in [m]$.

– Signal momentum: $\mathcal{G}_r^{(t)} := \langle g_r^{(t)}, w^* \rangle$ for $r \in [m]$.

– Max signal momentum: $\mathcal{G}^{(t)} = \mathcal{G}_{r_{\max}}^{(t)}$, where $r_{\max} = \mathrm{argmax}_{r \in [m]} c_r^{(t)}$.

– Noise momentum: $G_{i,j,r}^{(t)} = \langle g_r^{(t)}, X_i[j] \rangle$ for $i \in [N]$, $j \in [P] \backslash \{P(X_i)\}$ and $r \in [m]$.

– Max noise momentum: $G^{(t)} = \max_{r \in [m], j \in [P], j \neq P(X^{(i)})} G_{i,j,r}^{(t)}$, where $r_{\max} = \mathrm{argmax}_{r \in [m]} c_r^{(t)}$.

C Induction hypotheses

We prove our main result using an induction. More specifically, we make the following assumptions for every time $t \leq T$.

Induction hypothesis C.1 (Bound on the noise component for GD)**.** *Throughout the training process using GD for $t \leq T$, we maintain that:*

1. *(Large signal data have small noise component). For every $i \in \mathcal{Z}_1$, for every $j \in [P] \backslash \{P(X^{(i)})\}$ and $r \in [m]$, we maintain:*

$$|\Xi_{i,j,r}^{(t)}| \leq \tilde{O}(\sigma_0 \sigma \sqrt{d}). \tag{1}$$

2. *(Small signal data have large noise component). For every $i \in \mathcal{Z}_2$, for every $j \in [P] \backslash \{P(X^{(i)})\}$ and $r \in [m]$, we have:*

$$|\Xi_{i,j,r}^{(t)}| \leq \tilde{O}(1), \quad y_i \Xi_{i,j,r}^{(t)} \geq -\tilde{O}(\sigma_0 \sigma \sqrt{d}). \tag{2}$$

Induction hypothesis C.2 (Bound on the signal component for GD)**.** *Throughout the training process using GD for $t \leq T$, the signal component is bounded for every $r \in [m]$ as*

$$-\tilde{O}(\sigma_0) \leq c_r^{(t)} \leq \widetilde{O}(1/\alpha).$$

Induction hypothesis C.3 (Max noise is bounded by max signal component)**.** *Throughout the training process using GD for $t \leq T$, we maintain:*

$$\alpha \min\{\kappa, \alpha^2 (c^{(t)})^2\} \geq \tilde{\Omega}\left(\Xi_{\max}^{(t)}\right),$$

where $\kappa = \tilde{O}(1)$.

Induction hypothesis C.4 (Bound on the noise component for GD+M)**.** *Throughout the training process using GD+M for $t \leq T$, for every $i \in [N]$, for every $j \in [P] \backslash \{P(X^{(i)})\}$, we have that:*

$$|\Xi_{i,j,r}^{(t)}| \leq \tilde{O}(\sigma_0 \sigma \sqrt{d}) \tag{3}$$

In what follows, we aim at proving these hypotheses for $t = T + 1$.

D GRADIENTS

We start by computing the gradient of the loss $\widehat{L}$.

Lemma D.1 (Gradient of $\widehat{L}$). *For $t \geq 0$ and $r \in [m]$, the gradient of the loss $\widehat{L}$ with respect to w_r is:*

$$\nabla_{w_r} \widehat{L}(W^{(t)}) = -\frac{3}{N}\left[\left(\sum_{i\in\mathcal{Z}_1} \alpha^3 \ell_i^{(t)} + \sum_{i\in\mathcal{Z}_2} \beta^3 \ell_i^{(t)}\right)(c_r^{(t)})^2 w^* + \sum_{i=1}^{N} \sum_{j\neq P(X_i)} \ell_i^{(t)} (\Xi_{i,j,r}^{(t)})^2 X_i[j]\right].$$

Proof of Lemma D.1. . We derive $\widehat{L}$ with respect to w_r and obtain:

$$\nabla_{w_r} \widehat{L}(W^{(t)}) = -\frac{3}{N}\sum_{i=1}^{N}\sum_{j=1}^{P} \frac{y_i \langle w_r^{(t)}, X_i[j]\rangle^2}{1+\exp(f_{W^{(t)}}(X_i))} X_i[j]. \tag{4}$$

By rewriting equation (4), we obtain the desired result. □

To track the signal learnt by our models, we need the signal gradient which is the projection of the gradient $\nabla_{w_r}\widehat{L}$ on w^*.

Lemma D.2 (Signal gradient). *For all $t \geq 0$ and $r \in [m]$, the signal gradient is:*

$$-\mathscr{G}_r^{(t)} = \frac{3}{N}\left(\sum_{i\in\mathcal{Z}_1} \alpha^3 \ell_i^{(t)} + \sum_{i\in\mathcal{Z}_2} \beta^3 \ell_i^{(t)}\right)(c_r^{(t)})^2.$$

Proof of Lemma D.2. We obtain the desired result by projecting the gradient from Lemma D.1 on w^* and using $X_i[j] \perp w^*$. □

To prove the memorization of GD and the non-memorization of GD+M, we also need to compute the noise gradient which is the projection of the gradient $\nabla_{w_r}\widehat{L}$ on $X_i[j]$.

Lemma D.3 (Noise gradient). *For all $t \geq 0$, $i \in [N]$ and $j \in [P]\backslash\{P(X_i)\}$ and $r \in [m]$, the noise gradient is:*

$$\begin{aligned}
-G_{i,j,r}^{(t)} = &\frac{3}{N}\ell_i^{(t)}(\Xi_{i,j,r}^{(t)})^2 \|X_i[j]\|_2^2 \\
&+ \frac{3}{N}\sum_{k\neq P(X_i)} \ell_i^{(t)}(\Xi_{i,k,r}^{(t)})^2 \langle X_i[k], X_i[j]\rangle \\
&+ \frac{3}{N}\sum_{a\neq i}\sum_{k\neq P(X_a)} \ell_a^{(t)}(\Xi_{a,k,r}^{(t)})^2 \langle X_a[k], X_i[j]\rangle.
\end{aligned}$$

Proof of Lemma D.3. Similarly to Lemma D.2, we obtain the desired result by projecting the gradient from Lemma D.1 on $X_i[j]$ and using $X_i[j] \perp w^*$. □

Remark 1. *The gradient in Lemma D.1 involve sigmoid terms $\ell_i^{(t)}$. In several parts of the proof, we focus on the time where these terms are small. We consider that the sigmoid term is small for a κ such that*

$$\sum_{\tau=0}^{T} \frac{1}{1+\exp(\kappa)} \leq \tilde{O}(1) \implies \kappa \geq \log(\tilde{\Omega}(T)) \iff \kappa \geq \tilde{\Omega}(1). \tag{5}$$

Intuitively, equation (5) means that the sum of the sigmoid terms for all time steps is bounded (up to a logarithmic dependence).

E LEARNING WITH GD

In this section, we prove the lemmas in Section 4 and Theorem 3.1.

E.1 LEARNING SIGNAL WITH GD

To track the amount of signal learnt by GD, we make use of the following update.

Lemma E.1 (Signal update). *For all $t \geq 0$ and $r \in [m]$, the signal update equation (GD-S) is equal:*

$$c_r^{(t+1)} = c_r^{(t)} + 3\eta \left(\alpha^3 \nu_1^{(t)} + \beta^3 \nu_2^{(t)}\right) (c_r^{(t)})^2.$$

Consequently, it satisfies:

$$\tilde{\Theta}(\eta)(1-\hat{\mu})\alpha^3 \widehat{\ell}^{(t)}(\alpha)(c_r^{(t)})^2 \leq c_r^{(t+1)} - c_r^{(t)} \leq \tilde{\Theta}(\eta)\left((1-\hat{\mu})\alpha^3 \widehat{\ell}^{(t)}(\alpha) + \beta^3 \nu_2^{(t)}\right)(c_r^{(t)})^2.$$

Proof of Lemma E.1. The signal update is obtained by using equation (GD-S) and the signal gradient (Lemma D.2). This yields

$$c_r^{(t+1)} = c_r^{(t)} + \frac{3\eta}{N}\left(\sum_{i\in\mathcal{Z}_1} \alpha^3 \ell_i^{(t)} + \sum_{i\in\mathcal{Z}_2} \beta^3 \ell_i^{(t)}\right)(c_r^{(t)})^2. \tag{6}$$

To obtain the desired lower bound, we first drop the sum over $\mathcal{Z}_2$ in equation (6). Then, for $i \in \mathcal{Z}_1$, we apply Lemma H.1 to get $\ell_i^{(t)} = \tilde{\Theta}(1)\widehat{\ell}^{(t)}(\alpha)$.

To obtain the desired upper bound, we apply the same reasoning as above to bound the $\mathcal{Z}_1$ term. □

E.1.1 EARLY STAGES OF THE LEARNING PROCESS $t \in [0, T_0]$: LEARNING $\mathcal{Z}_1$ DATA

As we initialize $w_r^{(0)} \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I})$ with σ_0 small, the sigmoid terms $\widehat{\ell}^{(t)}(\alpha)$ and $\ell_i^{(t)}$ in the signal update are large at early iterations. As $c_r^{(t)}$ is non-decreasing (by Lemma E.1), $\widehat{\ell}^{(t)}(\alpha)$ eventually becomes small at a time $T_0 > 0$. As mentioned above, in this paper, we assume that the sigmoid term $\mathfrak{S}(x)$ becomes small when $x \geq \tilde{\Omega}(1)$. We therefore simplify equation (6) for $t \in [0, T_0]$.

Lemma E.2 (Signal update at early iterations). *Let $T_0 > 0$ the time where there exists $s \in [m]$ such that $c_s^{(t)} \geq \tilde{\Omega}(1/\alpha)$. Then, for $t \in [0, T_0]$ and for all $r \in [m]$, the signal update is simplified as:*

$$\tilde{\Theta}(\eta)(1-\hat{\mu})\alpha^3 (c_r^{(t)})^2 \leq c_r^{(t+1)} - c_r^{(t)} \leq \tilde{\Theta}(\eta)\left((1-\hat{\mu})\alpha^3 + \hat{\mu}\beta^3\right)(c_r^{(t)})^2. \tag{7}$$

Proof of Lemma E.2. For $t \in [0, T_0]$, we know that for all $s \in [m]$, we have $c_s^{(t)} \leq \frac{\tilde{O}(1)}{m^{1/3}\alpha} = \frac{\tilde{O}(1)}{\alpha}$ (since $m = \tilde{O}(1)$). Therefore, we have

$$\frac{1}{1+\exp(\tilde{\Omega}(1))} \leq \widehat{\ell}^{(t)}(\alpha) = \frac{1}{1+\exp\left(\sum_{s=1}^m \alpha^3 (c_s^{(t)})^3\right)} \leq 1. \tag{8}$$

By Remark 1, we know that the sigmoid is small only when we have $\frac{1}{1+\exp(\tilde{\Omega}(1))}$. From equation (8), we thus have:

$$\widehat{\ell}^{(t)}(\alpha) = \Theta(1). \tag{9}$$

Plugging equation (9) in the left-hand side of the inequality in Lemma E.1 yields the desired lower bound.

To obtain the desired upper bound, we first consider the upper bound from Lemma E.1. We upper bound $\frac{1}{N}\sum_{i\in\mathcal{Z}_2} \ell_i^{(t)} \leq \hat{\mu}$ since $\ell_i^{(t)} \leq 1$. Moreover, we use equation (9) to upper bound the $\widehat{\ell}^{(t)}(\alpha)$ term. □

We now prove Lemma 4.1 that quantifies the amount of signal learnt by GD when the gradient is large.

Lemma 4.1. *For all $r \in [m]$ and $t \geq 0$, equation (GD-S) is simplified as:*

$$c_r^{(t+1)} \geq c_r^{(t)} + \tilde{\Theta}(\eta)\alpha^3 (c_r^{(t)})^2 \cdot \mathrm{sigmoid}(-\textstyle\sum_{s=1}^t \alpha^3 (c_s^{(t)})^3).$$

Consequently, after $T_0 = \tilde{\Theta}\left(\frac{1}{\eta\alpha^3\sigma_0}\right)$ iterations, for all $t \in [T_0, T]$, we have $c^{(t)} \geq \tilde{\Omega}(1/\alpha)$.

Proof of Lemma 4.1. Let $r \in [m]$. By Lemma E.2, the signal update for $t \in [0, T_0]$ is

$$\begin{cases} c_r^{(t+1)} \leq c_r^{(t)} + A(c_r^{(t)})^2 \\ c_r^{(t+1)} \geq c_r^{(t)} + B(c_r^{(t)})^2 \end{cases}, \tag{10}$$

where A and B are respectively defined as:

$$A := \tilde{\Theta}(\eta)\left((1-\hat{\mu})\alpha^3 + \hat{\mu}\beta^3\right),$$
$$B := \tilde{\Theta}(\eta)(1-\hat{\mu})\alpha^3.$$

Now, we would like to find the time T_0 where $c_r^{(t)} \geq \tilde{\Omega}(1/\alpha)$. This time exists as $c_r^{(t)}$ is non-decreasing. To this end, we apply the Tensor Power method (Lemma J.15). This lemma only applies to non-negative sequences. Since we initialize the weights $w_r^{(0)} \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}_d)$, we have $c_r^{(0)} \sim \mathcal{N}(0, \sigma_0^2)$. Since all the $w_r^{(0)}$'s are i.i.d. so do the $c_r^{(0)}$'s. Therefore, the probability that at least one of the $c_r^{(0)}$ is non-negative is $1 - (1/2)^m = 1 - o(1)$. We thus conclude that with high probability, there exist an index $r \in [m]$ such that $c_r^{(0)} \geq 0$. Among the possible indices r that satisfy this inequality, we now focus on $r = r_{\max}$ where $r_{\max} \in \operatorname{argmax} c_r^{(0)}$.

Setting $\upsilon = \tilde{\Theta}(1/\alpha)$ in Lemma J.15, we deduce that the time t_0 is

$$t_0 = \frac{\tilde{\Theta}(1)}{\eta\alpha^3\sigma_0} + \frac{\tilde{\Theta}(1)\left((1-\hat{\mu})\alpha^3 + \hat{\mu}\beta^3\right)}{(1-\hat{\mu})\alpha^3}\left\lceil \frac{-\log(\tilde{\Theta}(\sigma_0\alpha))}{\log(2)} \right\rceil$$

□

We now turn to the proof of Lemma 4.2. It states that since the correlation of the GD iterates with the signal $c^{(t)}$ significantly increased, the $\mathcal{Z}_1$ gradients are now small. Before proving this result, we introduce an auxiliary Lemma.

Lemma E.3 (Lower bound on the signal update)**.** *Run GD on the loss function $\widehat{L}(W)$. After $T_0 = \tilde{\Theta}\left(\frac{1}{\eta\alpha^3\sigma_0}\right)$ iterations, the signal update is satisfies for $t \geq t_0$*

$$c_r^{(t+1)} \geq c_r^{(t)} + \eta\Omega(\alpha)\nu_1^{(t)}.$$

Proof of Lemma E.3 . From Lemma E.1, the signal update satisfies

$$c_r^{(t+1)} \geq c_r^{(t)} + \tilde{\Theta}(\eta)\nu_1^{(t)}\alpha^3(c_r^{(t)})^2. \tag{11}$$

We focus on equation (11) for the case where $r = r_{\max}$. Plugging $c^{(t)} \geq \tilde{\Omega}(1/\alpha)$ (Lemma 4.1) in equation (11), we obtain the desired result. □

Lemma 4.2. *Let $T_0 = \tilde{\Theta}\left(\frac{1}{\eta\alpha^3\sigma_0}\right)$. After $t \in [T_0, T]$ iterations, the $\mathcal{Z}_1$ derivative is bounded as $\nu_1^{(t)} \leq \tilde{O}\left(\frac{1}{\eta(t-T_0+1)\alpha}\right) + \tilde{O}\left(\frac{\beta^3}{\alpha}\right)\nu_2^{(t)}$. The full derivative is $\nu^{(t)} \leq \tilde{O}\left(\frac{1}{\eta(t-T_0+1)\alpha} + \left(1 + \frac{\beta^3}{\alpha}\right)\nu_2^{(t)}\right)$.*

Proof of Lemma 4.2. From Lemma E.3, we deduce an upper bound on $\nu_1^{(t)}$:

$$\nu_1^{(t)} \leq \tilde{O}\left(\frac{1}{\eta\alpha}\right)(c^{(t+1)} - c^{(t)}). \tag{12}$$

On the other hand, using Lemma D.2, the signal difference is bounded as:

$$\begin{aligned} c^{(t+1)} - c^{(t)} &\leq \sum_{r=1}^{m} c_r^{(t+1)} - c_r^{(t)} \\ &\leq (1-\hat{\mu})\Theta(\eta\alpha)\sum_{r=1}^{m}(\alpha c_r^{(t)})^2\widehat{\ell}^{(t)}(\alpha) + \hat{\mu}\Theta(\eta\beta^3)\sum_{r=1}^{m}(c_r^{(t)})^2\nu_2^{(t)}. \end{aligned} \tag{13}$$

By applying Induction hypothesis C.1 in equation (13) and using $m = \tilde{\Theta}(1)$, we obtain:

$$c^{(t+1)} - c^{(t)} \leq (1 - \hat{\mu})\Theta(\eta\alpha) \sum_{r=1}^{m} (\alpha c_r^{(t)})^2 \widehat{\ell}^{(t)}(\alpha) + \hat{\mu}\tilde{O}(\eta\beta^3)\nu_2^{(t)}. \tag{14}$$

We now bound equation (14) by a loss term by applying Lemma J.20. Using Lemma 4.1 and Induction hypothesis C.2, we have:

$$0 < \tilde{\Omega}(1/\alpha) \leq \tilde{\Omega}(1/\alpha) - m\tilde{O}(\sigma_0) \leq c^{(t)} - \sum_{r \neq r_{\max}} c_r^{(t)} \leq \sum_{r=1}^{m} \alpha c_r^{(t)} \leq m\tilde{O}(1) \leq \tilde{O}(1). \tag{15}$$

We can now apply Lemma J.20 and get:

$$\sum_{r=1}^{m} (\alpha c_r^{(t)})^2 \widehat{\ell}^{(t)}(\alpha) \leq \frac{20m\alpha e^{m\tilde{O}(\sigma_0)}}{\tilde{\Omega}(1)} \widehat{\mathcal{L}}^{(t)}(\alpha) \leq \tilde{O}(\alpha)\widehat{\mathcal{L}}^{(t)}(\alpha). \tag{16}$$

Plugging equation (16) in equation (14) yields:

$$c^{(t+1)} - c^{(t)} \leq (1 - \hat{\mu})\tilde{O}(\eta\alpha^2)\widehat{\mathcal{L}}^{(t)}(\alpha) + \hat{\mu}\tilde{O}(\eta\beta^3)\nu_2^{(t)}. \tag{17}$$

Combining equation (12) and equation (17), we thus obtain:

$$\nu_1^{(t)} \leq \tilde{O}\left(\frac{1}{\alpha}\right)\left((1 - \hat{\mu})\tilde{O}(\alpha^2)\widehat{\mathcal{L}}^{(t)}(\alpha) + \hat{\mu}\tilde{O}(\beta^3)\nu_2^{(t)}\right). \tag{18}$$

Lemma H.9 quantifies the convergence rate of the loss $\widehat{\mathcal{L}}^{(t)}(\alpha)$. We use it to obtain the desired bound on $\nu_1^{(t)}$.

The bound on $\nu^{(t)}$ is obtained by using its definition $\nu^{(t)} = \nu_1^{(t)} + \nu_2^{(t)}$. □

E.1.2 Late stages of learning process $t \in [T_0, T]$: amount of learnt signal controlled by $\mathcal{Z}_2$ derivative

We proved in the previous section that after T_0 iterations, the amount of signal learnt by the model with GD significantly increased until making the $\mathcal{Z}_1$ derivative small. We therefore need to rewrite the signal update in this case.

Lemma E.4 (Signal update at late iterations)**.** *For $t \in [T_0, T]$ and $r \in [m]$, the signal update equation (GD-S) satisfies:*

$$c_r^{(t+1)} - c_r^{(t)} = \tilde{\Theta}(\eta)\left(\alpha\nu_1^{(t)} \min\{\kappa, (c_r^{(t)})^2\alpha^2\} + \beta^3\nu_2^{(t)}(c_r^{(t)})^2\right).$$

Proof of Lemma E.4. From the signal update given by Lemma E.1, we know that:

$$c_r^{(t+1)} = c_r^{(t)} + \frac{3\eta}{N}\left(\sum_{i \in \mathcal{Z}_1} \alpha^3 \ell_i^{(t)} + \sum_{i \in \mathcal{Z}_2} \beta^3 \ell_i^{(t)}\right)(c_r^{(t)})^2. \tag{19}$$

To obtain the desired result, we first need to prove for $i \in \mathcal{Z}_1$:

$$\alpha^3 \ell_i^{(t)}(c^{(t)})^2 = \tilde{\Theta}(\alpha) \min\{\kappa, \alpha^2(c^{(t)})^2\}\ell_i^{(t)}. \tag{20}$$

We first remark that:

$$\alpha^3 \ell_i^{(t)}(c^{(t)})^2 = \frac{\alpha^3(c^{(t)})^2}{1 + \exp\left(\alpha^3 \sum_{s=1}^{m}(c_s^{(t)})^3 + \Xi_i^{(t)}\right)}. \tag{21}$$

By using Induction hypothesis C.1 and Induction hypothesis C.2, equation (21) is bounded as:

$$\begin{aligned}
\alpha^3 \ell_i^{(t)}(c^{(t)})^2 &= \frac{\tilde{\Theta}(\alpha^3)(c^{(t)})^2}{1 + \exp\left(\alpha^3(c^{(t)})^3 + \alpha^3 \sum_{s \neq r_{\max}}(c_s^{(t)})^3 + \Xi_i^{(t)}\right)} \\
&\leq \frac{\tilde{\Theta}(\alpha^3)(c^{(t)})^2}{1 + \exp\left(\alpha^3(c^{(t)})^3 - \tilde{O}(m\alpha^3\sigma_0^3) - \tilde{O}(mP(\sigma\sigma_0\sqrt{d})^3)\right)},
\end{aligned} \tag{22}$$

From Remark 1, we know that the sigmoid term in equation (22) becomes small when $\alpha^3(c^{(t)})^3 \geq \kappa$ which implies $\alpha c^{(t)} \geq \kappa^{1/3} = \tilde{\Omega}(1)$. To summarize, we have:

$$\alpha^3 \ell_i^{(t)} (c^{(t)})^2 = \begin{cases} \alpha^3 \ell_i^{(t)} (c^{(t)})^2 & \text{if } \alpha c^{(t)} \leq \tilde{O}(1) \\ 0 & \text{otherwise} \end{cases}. \tag{23}$$

Using equation (23), we therefore proved that $\alpha^3 (c^{(t)})^2 \ell_i^{(t)} = \alpha \min\{\kappa, (\alpha c^{(t)})^2\} \ell_i^{(t)}$.

Besides, we use Induction hypothesis C.2 to bound $(c_r^{(t)})^2$ in the right-hand side of the signal update. □

We now show that once the $\mathcal{Z}_1$ derivative is small, the amount of learnt signal is controlled by the $\mathcal{Z}_2$ derivative.

Lemma 4.3. *Let* $T_0 = \tilde{\Theta}\left(\frac{1}{\eta \alpha^3 \sigma_0}\right)$. *For* $t \in [T_0, T]$, *equation (GD-S) becomes* $c^{(t+1)} \leq \tilde{O}(1/\alpha) + \tilde{O}(\eta \beta^3 / \alpha) \sum_{\tau=T_0}^{t} \nu_2^{(\tau)}$.

Proof of Lemma 4.3. Let $\tau \in [T]$. From Lemma E.4, we know that:

$$c^{(\tau+1)} - c^{(\tau)} = \tilde{\Theta}(\eta) \left(\alpha \nu_1^{(\tau)} \min\{\kappa, (c^{(\tau)})^2 \alpha^2\} + \beta^3 \nu_2^{(\tau)} (c^{(\tau)})^2 \right) \tag{24}$$

By applying Induction hypothesis C.2 to bound $(c_r^{(t)})^2$ in equation (24), we obtain:

$$c^{(\tau+1)} - c^{(\tau)} \leq \tilde{\Theta}(\eta \alpha) \nu_1^{(\tau)} \min\{\kappa, (c^{(\tau)})^2 \alpha^2\} + \tilde{O}(\eta \beta^3 / \alpha^2) \nu_2^{(\tau)}. \tag{25}$$

Let $t \in [T]$. We now sum up equation (25) for $\tau = T_0, \ldots, t$ and obtain:

$$c^{(t+1)} \leq c^{(T_0)} + \tilde{\Theta}(\eta \alpha) \sum_{\tau=T_0}^{t} \nu_1^{(\tau)} \min\{\kappa, (c^{(\tau)})^2 \alpha^2\} + \tilde{O}(\eta \beta^3 / \alpha^2) \sum_{\tau=T_0}^{t} \nu_2^{(\tau)}. \tag{26}$$

We now plug the bound on $\nu_1^{(t)}$ from Lemma 4.2 in equation (26). This yields:

$$c^{(t+1)} \leq c^{(T_0)} + \sum_{\tau=T_0}^{t} \frac{\tilde{O}(1)}{\tau - T_0 + 1} + \tilde{O}(\eta \beta^3 / \alpha^2) \sum_{\tau=T_0}^{t} \nu_2^{(\tau)}. \tag{27}$$

Plugging $\sum_\tau 1/\tau \leq \tilde{O}(1)$ and $c^{(T_0)} \leq \tilde{O}(1/\alpha)$ (Induction hypothesis C.2) in equation (27), we obtain:

$$c^{(t+1)} \leq \tilde{O}(1/\alpha) + \tilde{O}(\eta \beta^3 / \alpha^2) \sum_{\tau=T_0}^{t} \nu_2^{(\tau)}.$$

□

E.2 Memorizing with GD

E.2.1 Early stages of memorization process $t \in [0, T_1]$: memorizing $\mathcal{Z}_2$ data

After Step 1, the gradient from Z_2 data dominates. We want to show that this leads the model to memorize. To this end, we first derive the noise update.

Lemma E.5 (Noise update)**.** *For all* $t \geq 0$, $i \in [N]$, $j \in [P] \backslash \{P(X_i)\}$ *and* $r \in [m]$, *the noise update equation (GD-N) is:*

$$\left| y_i \Xi_{i,j,r}^{(t+1)} - y_i \Xi_{i,j,r}^{(t)} - \frac{\tilde{\Theta}(\eta \sigma^2 d)}{N} \ell_i^{(t)} (\Xi_{i,j,r}^{(t)})^2 \right| \leq \frac{\tilde{\Theta}(\eta \sigma^2 \sqrt{d})}{N} \sum_{a=1}^{N} \ell_a^{(t)} \sum_{k \neq P(X_a)} (\Xi_{a,k,r}^{(t)})^2. \tag{N}$$

Proof of Lemma E.5. Let $i \in [N]$, $j \in [P]\backslash\{P(X_i)\}$ and $r \in [m]$. Using the definition of the gradient (Lemma D.1) and $X_i[j] \perp w^*$, equation (GD-N) becomes:

$$y_i\Xi_{i,j,r}^{(t+1)} = y_i\Xi_{i,j,r}^{(t)} + \frac{3\eta}{N}\ell_i^{(t)}(\Xi_{i,j,r}^{(t)})^2\|X_i[j]\|_2^2 + \frac{3\eta}{N}\ell_i^{(t)}\sum_{\substack{k\neq P(X_i)\\k\neq j}}(\Xi_{i,k,r}^{(t)})^2\langle X_i[k], X_i[j]\rangle + \frac{3\eta}{N}\sum_{a\neq i}\ell_a^{(t)}\sum_{k\neq P(X_a)}(\Xi_{a,k,r}^{(t)})^2\langle X_a[k], X_i[j]\rangle. \tag{28}$$

We now use Lemma J.5 and Lemma J.7 to respectively bound $\|X_i[j]\|_2^2$ and $\langle X_a[k], X_i[j]\rangle$ in equation (28) and obtain the desired result. □

In the next lemma, we further simplify the noise update from Lemma E.5.

Lemma E.6 (Sum of noise updates)**.** *Let $i \in \mathcal{Z}_2$, $j \in [P]\backslash\{P(X_i)\}$ and $r \in [m]$. Let $\mathfrak{T} = \tilde{\Theta}\left(\frac{P\sigma^2\sqrt{d}}{\eta\beta^3\hat{\mu}}\right)$. For $t \leq \mathfrak{T}$, the noise update satisfies:*

$$\left|y_i\Xi_{i,j,r}^{(t+1)} - y_i\Xi_{i,j,r}^{(0)} - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N}\sum_{\tau=0}^{t}\widehat{\ell}^{(\tau)}(\Xi_i^{(\tau)})(\Xi_{i,j,r}^{(\tau)})^2\right| \leq \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right). \tag{N-I}$$

Proof of Lemma E.6. Lemma H.5 is the main lemma we use to derive noise updates in the analysis. For $i \in \mathcal{Z}_2$, $j \in [P]\backslash\{P(X_i)\}$ and $r \in [m]$, it states that:

$$\left|y_i\Xi_{i,j,r}^{(t)} - y_i\Xi_{i,j,r}^{(0)} - \frac{\eta\tilde{\Theta}(\sigma^2 d)}{N}\sum_{\tau=0}^{t-1}\ell_i^{(\tau)}(\Xi_{i,j,r}^{(\tau)})^2\right| \leq \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right) + \tilde{O}\left(\frac{\eta\beta^3}{\alpha}\right)\sum_{j=0}^{t}\nu_2^{(j)}. \tag{29}$$

Since $t \leq \mathfrak{T} = \tilde{\Theta}\left(\frac{P\sigma^2\sqrt{d}}{\eta\beta^3\hat{\mu}}\right)$, we bound the second sum term in equation (29) as:

$$\tilde{O}\left(\frac{\eta\beta^3}{\alpha N}\right)\sum_{j=0}^{t}\nu_2^{(j)} \leq \tilde{O}\left(\frac{\eta\beta^3}{\alpha N}\right)\hat{\mu}t \leq \tilde{O}\left(\frac{\eta\beta^3\hat{\mu}\mathfrak{T}}{\alpha N}\right) \leq \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right). \tag{30}$$

From equation (30), we deduce that

$$\left|y_i\Xi_{i,j,r}^{(t)} - y_i\Xi_{i,j,r}^{(0)} - \frac{\eta\tilde{\Theta}(\sigma^2 d)}{N}\sum_{\tau=0}^{t-1}\ell_i^{(\tau)}(\Xi_{i,j,r}^{(\tau)})^2\right| \leq \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right). \tag{31}$$

By applying Lemma H.2, we have $\ell_i^{(\tau)} = \tilde{\Theta}(1)\widehat{\ell}^{(t)}(\Xi_i^{(\tau)})$. Plugging this in equation (31) yields the desired result.

□

Lemma 4.4. *Let $t \geq 0$ and $i \in \mathcal{Z}_2$. Assume that $\Xi_i^{(t)} \leq \tilde{O}(1)$. Then, equation (GD-N) can be simplified as:*

$$y_i\Xi_{i,j,r}^{(t+1)} \geq y_i\Xi_{i,j,r}^{(0)} + \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N}\sum_{\tau=0}^{t}(\Xi_{i,j,r}^{(\tau)})^2 - \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right).$$

Let $T_1 = \tilde{O}\left(\frac{N}{\sigma_0\sigma\sqrt{d}\sigma^2 d}\right)$. Therefore, $\Xi_i^{(t)} \geq \tilde{\Omega}(1)$, for $t \in [T_1, T]$ and thus GD memorizes.

Proof of Lemma 4.4. Let $t \leq \mathfrak{T}$, where $\mathfrak{T} = \tilde{\Theta}\left(\frac{P\sigma^2\sqrt{d}}{\eta\beta^3\hat{\mu}}\right)$. Let $i \in \mathcal{Z}_2$, $j \in [P]\backslash\{P(X_i)\}$ and $r \in [m]$. From Lemma E.6, we know that

$$\left|y_i\Xi_{i,j,r}^{(t+1)} - y_i\Xi_{i,j,r}^{(0)} - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N}\sum_{\tau=0}^{t}\widehat{\ell}^{(\tau)}(\Xi_i^{(\tau)})(\Xi_{i,j,r}^{(\tau)})^2\right| \leq \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right). \tag{32}$$

In particular, equation (32) implies that:

$$y_i \Xi_{i,j,r}^{(t+1)} \geq y_i \Xi_{i,j,r}^{(0)} + \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N} \sum_{\tau=0}^{t} \widehat{\ell}^{(\tau)}(\Xi_i^{(\tau)})(\Xi_{i,j,r}^{(\tau)})^2 - \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right). \tag{33}$$

From Remark 1, we know that $\widehat{\ell}^{(\tau)}(\Xi_i^{(\tau)})$ is small when $\Xi_i^{(\tau)} \geq \kappa = \tilde{\Omega}(1)$. To have this condition, it is sufficient that there exist j, r such that $y_i \Xi_{i,j,r} \geq \tilde{\Omega}(1)$. Indeed, by using Induction hypothesis C.1, we see that:

$$\Xi_i^{(t)} = (y_i \Xi_{i,j,r}^{(t)})^3 + \sum_{s=1}^{m} \sum_{k \neq P(X_i)} (y_i \Xi_{i,k,s}^{(t)})^3 \geq \tilde{\Omega}(1) - \tilde{O}(mP(\sigma\sigma_0\sqrt{d})^3) \geq \tilde{\Omega}(1).$$

Therefore, when for every $j \in [P] \backslash \{P(X_i)\}$ and $r \in [m]$ we have $y_i \Xi_{i,j,r}^{(t)} \leq \tilde{O}(1)$, the noise update equation (32) becomes:

$$\left| y_i \Xi_{i,j,r}^{(t+1)} - y_i \Xi_{i,j,r}^{(0)} - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N} \sum_{\tau=0}^{t} (\Xi_{i,j,r}^{(\tau)})^2 \right| \leq \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right). \tag{34}$$

Using the recursion equation (34), we want to determine T_1, the time such that there exists j, r such that $y_i \Xi_{i,j,r} \geq \tilde{\Omega}(1)$. Let's assume for now that $T_1 \leq \mathfrak{T}$ (otherwise we cannot use equation (34)). We verify this condition at the end of the proof.

equation (34) indicates that the noise iterate satisfies for $t \in [0, T_1]$:

$$\begin{cases} y_i \Xi_{i,j,r}^{(t)} \geq y_i \Xi_{i,j,r}^{(0)} + A \sum_{\tau=0}^{t-1} (\Xi_{i,j,r}^{(\tau)})^2 - C \\ y_i \Xi_{i,j,r}^{(t)} \leq y_i \Xi_{i,j,r}^{(0)} + A \sum_{\tau=0}^{t-1} (\Xi_{i,j,r}^{(\tau)})^2 + C \end{cases}, \tag{35}$$

where $A, C > 0$ are constants defined as

$$A = \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N}, \quad C = \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right). \tag{36}$$

To find T_1, we apply the Tensor Power method (Lemma J.16) to equation (35). We initialize the weights $w_r^{(0)} \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}_d)$ and $X_i[j] \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$. Therefore, we have $\mathbb{P}[y_i \Xi_{i,j,r}^{(0)} \geq 0] = 1/2$. Since all the $w_r^{(0)}$'s are i.i.d. so do the $\Xi_{i,j,r}^{(0)}$'s. Therefore, the probability that at least one of the $\Xi_{i,j,r}^{(0)}$ is non-negative is $1 - (1/2)^m = 1 - o(1)$. We thus conclude that with high probability, there exist an index $r \in [m]$ such that $y_i \Xi_{i,j,r}^{(0)} \geq \Omega(\sigma\sigma_0\sqrt{d}) = \omega(C)$. In what follows, we focus on such index r.

Setting the constants A, C as in equation (36) and $\upsilon = \tilde{O}(1)$, the time T_1 obtained with the Tensor Power method is for $\delta \in (0, 1)$

$$T_1 = \frac{N(\delta+1)}{\tilde{\Theta}(\eta\sigma^2 d) y_i \Xi_{i,j,r}^{(0)}} + \frac{\tilde{\Theta}(\eta\sigma^2 d) + \tilde{O}\left(\frac{NP\sigma^2\sqrt{d}}{\alpha}\right)}{\tilde{\Theta}(\eta\sigma^2 d)(y_i \Xi_{i,j,r}^{(0)})^2} \left\lceil \frac{\log\left(\frac{\tilde{O}(1)}{y_i \Xi_{i,j,r}^{(0)}}\right)}{\log(1+\delta)} \right\rceil.$$

Since $y_i \Xi_{i,j,r}^{(0)} = \tilde{\Theta}(\sigma_0\sigma\sqrt{d})$, we obtain $T_1 = \tilde{O}\left(\frac{N}{\eta\sigma^2 d\sigma_0\sigma\sqrt{d}}\right)$. We indeed verify that $T_1 \leq \mathfrak{T}$ as $\tilde{O}\left(\frac{N}{\eta\sigma^2 d\sigma_0\sigma\sqrt{d}}\right) \ll \tilde{\Theta}\left(\frac{NP\sigma^2\sqrt{d}}{\eta\beta^3\tilde{\mu}}\right)$. □

E.2.2 Late stages of memorization $t \in [T_1, T]$: convergence to a minimum

We proved in the previous section that after T_1 iterations, the amount of noise memorized by the GD model significantly increased. We want to show that after this phase, the $\mathcal{Z}_2$ derivative converges to zero.

Lemma E.7 (Bound on $\mathcal{Z}_2$ derivative at late iterations). *Let* $T_1 = \tilde{O}\left(\frac{N}{\sigma_0\sigma\sqrt{d}\sigma^2 d}\right)$. *For* $t \in [T_1, T]$, *we have* $\sum_{\tau=T_1}^{t} \nu_2^{(\tau)} \leq \tilde{O}\left(\frac{1}{\eta\sigma_0}\right)$.

Proof of Lemma E.7. In Lemma 4.4, we proved that after T_1 iterations, for all $i \in \mathcal{Z}_2$ and $j \in [P]\backslash\{P(X_i)\}$, there exists $r \in [m]$ such that $y_i \Xi_{i,j,r}^{(t)} \geq \tilde{\Omega}(1)$. Therefore, for $t \in [T_1, T]$, there exists $r \in [m]$ such that the noise update (from Lemma H.5) satisfies:

$$\sum_{\tau=T_1}^{t} \nu_2^{(\tau)} \leq \tilde{O}\left(\frac{1}{\eta\sigma^2 d}\right) \sum_{i\in\mathcal{Z}_2} y_i(\Xi_{i,j,r}^{(t+1)} - \Xi_{i,j,r}^{(T_1)}) + \tilde{O}\left(\frac{P}{\alpha\eta\sqrt{d}}\right) + \tilde{O}\left(\frac{\beta^3}{\alpha\sigma^2 d}\right) \sum_{j=T_1}^{t-1} \nu_2^{(j)}. \tag{37}$$

On the other hand, from Lemma H.5, we know that for all $r \in [m]$:

$$\sum_{i\in\mathcal{Z}_2} y_i(\Xi_{i,j,r}^{(t+1)} - \Xi_{i,j,r}^{(T_1)}) \leq \frac{\eta\tilde{\Theta}(\sigma^2 d)}{N} \sum_{\tau=T_1}^{t-1} \sum_{i\in\mathcal{Z}_2} \ell_i^{(\tau)} (\Xi_{i,j,r}^{(\tau)})^2 + \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right) + \tilde{O}\left(\frac{\eta\beta^3}{\alpha}\right) \sum_{j=T_1}^{t-1} \nu_2^{(j)}. \tag{38}$$

Combining equation (37) and equation (38) yields:

$$\sum_{\tau=T_1}^{t} \nu_2^{(\tau)} \leq \frac{\tilde{O}(1)}{N} \sum_{\tau=T_1}^{t-1} \sum_{i\in\mathcal{Z}_2} \ell_i^{(\tau)} (\Xi_{i,j,r}^{(\tau)})^2 + \tilde{O}\left(\frac{\beta^3}{\alpha\sigma^2 d}\right) \sum_{j=T_1}^{t-1} \nu_2^{(j)} + \tilde{O}\left(\frac{P}{\eta\alpha\sqrt{d}}\right) \tag{39}$$

Again, because $\tilde{O}\left(\frac{\beta^3}{\alpha\sigma^2 d}\right) \ll 1$, we further simplify equation (39):

$$\sum_{\tau=T_1}^{t} \nu_2^{(\tau)} \leq \frac{\tilde{O}(1)}{N} \sum_{\tau=T_1}^{t-1} \sum_{i\in\mathcal{Z}_2} \ell_i^{(\tau)} (\Xi_{i,j,r}^{(\tau)})^2 + \tilde{O}\left(\frac{P}{\eta\alpha\sqrt{d}}\right). \tag{40}$$

By applying Lemma H.2 to bound $\ell_i^{(\tau)}$ on the right-hand side, equation (40) becomes:

$$\sum_{\tau=T_1}^{t} \nu_2^{(\tau)} \leq \frac{\tilde{O}(1)}{N} \sum_{\tau=T_1}^{t-1} \sum_{i\in\mathcal{Z}_2} \widehat{\ell}^{(\tau)}(\Xi_i^{(\tau)}) (\Xi_{i,j,r}^{(\tau)})^2 + \tilde{O}\left(\frac{P}{\eta\alpha\sqrt{d}}\right). \tag{41}$$

Since equation (41) holds for every $r \in [m]$ and $j \in [P]\backslash\{P(X_i)\}$, we sum it up and obtain:

$$\frac{1}{N} \sum_{\tau=T_1}^{t} \sum_{i\in\mathcal{Z}_2} \ell_i^{(\tau)} \leq \frac{\tilde{O}(1)}{Nm(P-1)} \sum_{\tau=T_1}^{t-1} \sum_{i\in\mathcal{Z}_2} \sum_{r=1}^{m} \sum_{j\neq P(X_i)} \widehat{\ell}^{(\tau)}(\Xi_i^{(\tau)}) (\Xi_{i,j,r}^{(\tau)})^2 + \tilde{O}\left(\frac{P}{\eta\alpha\sqrt{d}}\right). \tag{42}$$

Remark that $\frac{1}{m(P-1)} \leq \tilde{O}(1)$. Moreover, by applying Lemma J.21 to equation (42), we have:

$$\frac{1}{N} \sum_{\tau=T_1}^{t} \sum_{i\in\mathcal{Z}_2} \ell_i^{(\tau)} \leq \frac{\tilde{O}(1)}{N} \sum_{\tau=T_1}^{t-1} \sum_{i\in\mathcal{Z}_2} \widehat{\mathcal{L}}^{(\tau)}(\Xi_i^{(\tau)}) + \tilde{O}\left(\frac{P}{\eta\alpha\sqrt{d}}\right). \tag{43}$$

We now apply Lemma H.10 to bound the loss in equation (43).

$$\frac{1}{N} \sum_{\tau=T_1}^{t} \sum_{i\in\mathcal{Z}_2} \ell_i^{(\tau)} \leq \frac{\tilde{O}(1)}{\eta} \sum_{\tau=T_1}^{t} \frac{1}{\tau - T_1 + 1} + \tilde{O}\left(\frac{P}{\alpha\sqrt{d}}\right) \leq \tilde{O}\left(\frac{1}{\eta}\right) + \tilde{O}\left(\frac{P}{\eta\alpha\sqrt{d}}\right) \leq \tilde{O}\left(\frac{1}{\eta}\right), \tag{44}$$

where we used in equation (44) $\sum_{\tau=T_1+1}^{t} 1/\tau \leq \tilde{O}(1)$ and $P/\alpha = \tilde{O}(1)$.

□

Using Lemma E.7, we can obtain a bound on the sum over time of $\mathcal{Z}_2$ derivatives .

Lemma 4.5. *Let* $T_1 = \tilde{O}\left(\frac{N}{\sigma_0 \sigma \sqrt{d} \sigma^2 d}\right)$. *For* $t \in [T_1, T]$, *we have* $\sum_{\tau=0}^{t} \nu_2^{(\tau)} \leq \tilde{O}\left(\frac{1}{\eta \sigma_0}\right)$.

Proof of Lemma 4.5 . We know that:

$$\sum_{j=0}^{T_1-1} \nu_2^{(j)} \leq \hat{\mu} T_1. \tag{45}$$

Combining the bound on $\sum_{j=T_1}^{T} \nu_2^{(j)}$ from Lemma E.7 and equation (45) yields:

$$\sum_{j=0}^{T} \nu_2^{(j)} = \sum_{j=0}^{T_1-1} \nu_2^{(j)} + \sum_{j=T_1}^{T} \nu_2^{(j)} \leq \tilde{\Theta}\left(\frac{\hat{\mu} N}{\eta \sigma_0 \sigma \sqrt{d} \sigma^2 d}\right) + \tilde{O}\left(\frac{1}{\eta}\right) \leq \tilde{O}\left(\frac{1}{\eta \sigma_0}\right). \tag{46}$$

□

We have thus a control on the sum of $\mathcal{Z}_2$ derivatives. We can make use of Lemma 4.3 to get the final control on the signal iterate $c^{(t)}$.

Lemma 4.6. *For* $t \leq T$, *the signal component satisfies* $c^{(t)} \leq \tilde{O}(1/\alpha)$.

Proof of Lemma 4.6. Let $t \in [T]$. From Lemma 4.3, we know that the signal is bounded as

$$c^{(t)} \leq \tilde{O}(1/\alpha) + \tilde{O}(\eta \beta^3 / \alpha^2) \sum_{\tau=T_0}^{t-1} \nu_2^{(\tau)}. \tag{47}$$

We plug the bound from Lemma 4.5 to bound the last term in the right-hand side of equation (47).

□

We proved that the weights learnt by GD satisfy for $r \in [m]$

$$w_r^{(T)} = c_r^{(T)} w^* + v_r^{(T)}, \tag{48}$$

where for all $r \in [m]$, $c_r^{(T)} \leq \tilde{O}(1)$ (Lemma 4.6) and $v_r^{(T)} \in \text{span}(X_i[j]) \subset \text{span}(w^*)^{\perp}$. By Lemma 4.4, since $\Xi_i^{(t)} \geq \tilde{\Omega}(1)$, we have $\|v_r^{(T)}\|_2 \geq 1$. We are now ready to prove the generalization achieved by GD and stated in Theorem 3.1.

Theorem 3.1. *Assume that we run GD on (P) for* T *iterations with parameters set as in Parametrization 2.1. With probability at least* $1 - o(1)$, *the weights learned by GD*

1. *partially learn the feature: for all* $r \in [m]$, $|c_r^{(T)}| \leq \tilde{O}(1/\alpha)$.
2. *memorize from small margin data: for all* $i \in \mathcal{Z}_2$, $\Xi_i^{(t)} \geq \tilde{\Omega}(1)$.

Consequently, the training error is smaller than $O(\mu/\text{poly}(d))$ *and the test error is* ***at least*** $\tilde{\Omega}(\mu)$.

Proof of Theorem 3.1. We now bound the training and test error achieved by GD at time T.

Train error. Lemma H.10 provides a convergence bound on the training loss.

$$\widehat{L}(W^{(T)}) \leq \frac{\tilde{\Theta}(1)}{\eta(T - T_0 + 1)}. \tag{49}$$

Plugging $T \geq \text{poly}(d)N/\eta$ and $\mu = \Theta(1/N)$ in equation (49) yields:

$$\widehat{L}(W^{(T)}) \leq \tilde{O}\left(\frac{1}{\text{poly}(d)N}\right) \leq \tilde{O}\left(\frac{\mu}{\text{poly}(d)}\right). \tag{50}$$

Test error. Let (X, y) be a datapoint. We remind that $X = (X[1], \dots, X[P])$ where $X[P(X)] = \theta y w^*$ and $X[j] \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ for $j \in [P] \backslash \{P(X)\}$. We bound the test error as follows:

$$\begin{aligned}\mathscr{L}(f_{W^{(T)}}) &= \mathbb{E}_{\substack{\mathcal{Z} \sim \mathcal{D} \\ (X,y) \sim \mathcal{Z}}}[\mathbf{1}_{y f_{W^{(T)}}(X) < 0}] \\ &= \mathbb{E}_{(X,y)\sim \mathcal{Z}_1}[\mathbf{1}_{y f_{W^{(T)}}(X) < 0}]\mathbb{P}[\mathcal{Z}_1] + \mathbb{E}_{(X,y)\sim \mathcal{Z}_2}[\mathbf{1}_{y f_{W^T}(X) < 0}]\mathbb{P}[\mathcal{Z}_2] \\ &= (1-\hat{\mu})\mathbb{P}[y f_{W^{(T)}}((X) < 0 | (X,y) \sim \mathcal{Z}_1] + \hat{\mu}\mathbb{P}[y f_{W^{(T)}}(X) < 0 | (X,y) \sim \mathcal{Z}_2].\end{aligned} \tag{51}$$

We now want to compute the probability terms in equation (51). We remind that $(X, y) \sim \mathcal{Z}_1$, $y f_{W^{(T)}}(X)$ is given by

$$\begin{aligned} y f_{W^{(T)}}(X) &= y \sum_{s=1}^{m} \sum_{j=1}^{P} \langle w_s^{(T)}, X[j] \rangle^3 \\ &= \alpha^3 \sum_{s=1}^{m} (c_s^{(T)})^3 + y \sum_{s=1}^{m} \sum_{j \neq P(X)} \langle v_s^{(T)}, X[j] \rangle^3. \end{aligned} \tag{52}$$

We now apply Lemma 4.6 to equation (52) and obtain:

$$y f_{W^{(T)}}(X) \leq \tilde{O}(1) + y \sum_{s=1}^{m} \sum_{j \neq P(X)} \langle v_s^{(T)}, X[j] \rangle^3. \tag{53}$$

Let $(X, y) \sim \mathcal{Z}_2$. Similarly, by applying Lemma 4.6, $y f_{W^{(T)}}(X)$ is bounded as:

$$y f_{W^{(T)}}(X) \leq \tilde{O}((\beta/\alpha)^3) + + y \sum_{s=1}^{m} \sum_{j \neq P(X)} \langle v_s^{(T)}, X[j] \rangle^3. \tag{54}$$

Therefore, using equation (113), we upper bound the test error equation (111) as:

$$\begin{aligned} \mathscr{L}(f_{W^{(T)}}) \geq (1-\hat{\mu})\mathbb{P}\left[y \sum_{s=1}^{m} \sum_{j \neq P(X)} \langle v_s^{(T)}, X[j] \rangle^3 \leq -\tilde{\Omega}(1) \right] \\ + \hat{\mu}\mathbb{P}\left[y \sum_{s=1}^{m} \sum_{j \neq P(X)} \langle v_s^{(T)}, X[j] \rangle^3 \leq -\tilde{\Omega}((\beta/\alpha)^3) \right] \\ \geq \hat{\mu}\mathbb{P}\left[y \sum_{s=1}^{m} \sum_{j \neq P(X)} \langle v_s^{(T)}, X[j] \rangle^3 \leq -\tilde{\Omega}((\beta/\alpha)^3) \right]. \end{aligned} \tag{55}$$

Since y is taken uniformly from $\{-1, 1\}$, we further simplify equation (55) as:

$$\mathscr{L}(f_{W^{(T)}}) \geq \frac{\hat{\mu}}{2}\mathbb{P}\left[\left| \sum_{s=1}^{m} \sum_{j \neq P(X)} \langle v_s^{(T)}, X[j] \rangle^3 \right| \geq \tilde{\Omega}((\beta/\alpha)^3) \right]. \tag{56}$$

We know that $\tilde{\Theta}(\beta^3) = \tilde{\Theta}(\sigma^3)$. Therefore, we now apply Lemma J.12 to bound equation (56) and finally obtain:

$$\mathscr{L}(f_{W^{(T)}}) \geq \frac{\hat{\mu}}{2}\mathbb{P}\left[\left| \sum_{s=1}^{m} \sum_{j \neq P(X)} \langle v_s^{(T)}, X[j] \rangle^3 \right| \geq \tilde{\Omega}((\beta/\alpha)^3) \right] \geq \frac{\hat{\mu}}{2}\left(1 - \frac{\tilde{O}(d)}{2^d}\right) \geq \tilde{\Omega}(\mu). \tag{57}$$

□

E.3 Proof of the induction hypotheses for $t+1$

To prove Theorem 3.1, we used the induction hypotheses stated in Appendix C. The goal of this section is to prove them for $t+1$.

Proof of Induction hypothesis C.1. We prove here the main hypotheses we made on the noise when using GD.

GD Noise for $i \in \mathcal{Z}_2$. Let $i \in \mathcal{Z}_2$. We know that for $t \in [T]$, $y_i \Xi_{i,j,r}^{(t)} \leq \tilde{O}(1)$. Let's prove the result for $t+1$. From Lemma H.6, we have:

$$\left| y_i(\Xi_{i,j,r}^{(t+1)} - \Xi_{i,j,r}^{(0)}) - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N} \sum_{\tau=0}^{t} \ell_i^{(\tau)} \min\{\kappa, (\Xi_{i,j,r}^{(\tau)})^2\} \right| \leq \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right) + \tilde{O}\left(\frac{\eta\beta^3}{\alpha^2}\right) \sum_{\tau=0}^{t} \nu_2^{(\tau)}. \tag{58}$$

Let's start with the upper bound $y_i \Xi_{i,j,r}^{(t)}$ for $i \in \mathcal{Z}_2$. Using Lemma H.7, Lemma 4.5 and Induction hypothesis C.1, we deduce from equation (58) that:

$$y_i \Xi_{i,j,r}^{(t+1)} \leq \tilde{O}(1) + \tilde{O}(\sigma^2 d) + \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right) + \tilde{O}\left(\frac{\beta^3}{\alpha^2\sigma_0}\right) \leq \tilde{O}(1), \tag{59}$$

which proves the induction hypothesis for $t+1$. Regarding the lower bound, using Induction hypothesis C.1 and Lemma 4.5, we deduce from equation (58) that:

$$y_i \Xi_{i,j,r}^{(t+1)} \geq -\tilde{O}(\sigma\sigma_0\sqrt{d}) - \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right) - \tilde{O}\left(\frac{\beta^3}{\alpha^2\sigma_0}\right) \geq -\tilde{O}(\sigma\sigma_0\sqrt{d}), \tag{60}$$

which proves the induction hypothesis for $t+1$.

GD Noise for $i \in \mathcal{Z}_1$. Let $i \in \mathcal{Z}_1$. We know that for $t \in [T]$, $y_i \Xi_{i,j,r}^{(t)} \leq \tilde{O}(\sigma\sigma_0\sqrt{d})$. Let's prove the result for $t+1$. Using Lemma D.3, we know that the equation (GD-N) update is:

$$\begin{aligned} y_i \Xi_{i,j,r}^{(t+1)} \leq\ & y_i \Xi_{i,j,r}^{(0)} + \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N} \sum_{\tau=0}^{t} \ell_i^{(\tau)} (\Xi_{i,j,r}^{(\tau)})^2 \\ & + \frac{\tilde{\Theta}(\eta\sigma^2\sqrt{d})}{N} \sum_{a\in\mathcal{Z}_1} \sum_{k\neq P(X_a)} \sum_{\tau=0}^{t} \ell_a^{(\tau)} (\Xi_{a,k,r}^{(\tau)})^2 \\ & + \frac{\tilde{\Theta}(\eta\sigma^2\sqrt{d})}{N} \sum_{a\in\mathcal{Z}_2} \sum_{k\neq P(X_a)} \sum_{\tau=0}^{t} \ell_a^{(\tau)} (\Xi_{a,k,r}^{(\tau)})^2. \end{aligned} \tag{61}$$

Using Induction hypothesis C.1, we bound $y_i \Xi_{i,j,r}^{(0)}$ and $(\Xi_{a,k,r}^{(\tau)})^2$ in equation (61). We obtain:

$$\begin{aligned} y_i \Xi_{i,j,r}^{(t+1)} \leq\ & \tilde{O}(\sigma\sigma_0\sqrt{d}) + \frac{\tilde{\Theta}(\eta\sigma_0^2\sigma^4 d^2)}{N} \sum_{\tau=0}^{t} \ell_i^{(\tau)} \\ & + \tilde{\Theta}(\eta P\sigma_0^2\sigma^4 d^{3/2}) \sum_{\tau=0}^{t} \nu_1^{(\tau)} \\ & + \frac{\tilde{\Theta}(\eta\sigma^2\sqrt{d})}{N} \sum_{a\in\mathcal{Z}_2} \sum_{k\neq P(X_a)} \sum_{\tau=0}^{t} \ell_a^{(\tau)} (\Xi_{a,k,r}^{(\tau)})^2 \end{aligned} \tag{62}$$

Now, we apply Lemma H.4 to bound $\nu_1^{(\tau)}$ and $\ell_i^{(\tau)}/N$ in equation (62).

$$\begin{aligned} y_i \Xi_{i,j,r}^{(t+1)} \leq\ & \tilde{O}(\sigma\sigma_0\sqrt{d}) + \tilde{O}\left(\frac{\sigma_0^2\sigma^4 d^2}{\alpha}\right) + \tilde{O}\left(\frac{\eta\sigma_0^2\sigma^4 d^2\beta^3}{\alpha}\right) \sum_{\tau=0}^{t} \nu_2^{(\tau)} \\ & + \tilde{O}\left(\frac{P\sigma_0^2\sigma^4 d^{3/2}}{\alpha}\right) + \tilde{O}\left(\frac{\eta P\sigma_0^2\sigma^4 d^{3/2}\beta^3}{\alpha}\right) \sum_{\tau=0}^{t} \nu_2^{(\tau)} \\ & + \frac{\tilde{\Theta}(\eta\sigma^2\sqrt{d})}{N} \sum_{a\in\mathcal{Z}_2} \sum_{k\neq P(X_a)} \sum_{\tau=0}^{t} \ell_a^{(\tau)} (\Xi_{a,k,r}^{(\tau)})^2 \end{aligned} \tag{63}$$

We now apply Lemma H.6 and Lemma 4.5 to bound the derivative terms in equation (63).

$$\begin{aligned} y_i \Xi_{i,j,r}^{(t+1)} &\leq \tilde{O}(\sigma\sigma_0\sqrt{d}) + \tilde{O}\left(\frac{\sigma_0^2\sigma^4 d^2}{\alpha}\right) + \tilde{O}\left(\frac{\sigma_0\sigma^4 d^2\beta^3}{\alpha}\right) \\ &+ \tilde{O}\left(\frac{P\sigma_0^2\sigma^4 d^{3/2}}{\alpha}\right) + \tilde{O}\left(\frac{P\sigma_0\sigma^4 d^{3/2}\beta^3}{\alpha}\right) \\ &+ \tilde{O}(P\sigma^2\sqrt{d}) \\ &\leq \tilde{O}(\sigma\sigma_0\sqrt{d}), \end{aligned}$$

which proves the induction hypothesis for $t+1$. Now, let's prove that $y_i \Xi_{i,j,r}^{(t+1)} \geq -\tilde{O}(\sigma\sigma_0\sqrt{d})$. Similarly to above, the equation (GD-N) update is bounded as:

$$\begin{aligned} y_i \Xi_{i,j,r}^{(t+1)} &\geq y_i \Xi_{i,j,r}^{(0)} + \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N} \sum_{\tau=0}^{t} \ell_i^{(\tau)} (\Xi_{i,j,r}^{(\tau)})^2 \\ &- \frac{\tilde{\Theta}(\eta\sigma^2\sqrt{d})}{N} \sum_{a\in\mathcal{Z}_1} \sum_{k\neq P(X_a)} \sum_{\tau=0}^{t} \ell_a^{(\tau)} (\Xi_{a,k,r}^{(\tau)})^2 \\ &- \frac{\tilde{\Theta}(\eta\sigma^2\sqrt{d})}{N} \sum_{a\in\mathcal{Z}_2} \sum_{k\neq P(X_a)} \sum_{\tau=0}^{t} \ell_a^{(\tau)} (\Xi_{a,k,r}^{(\tau)})^2. \end{aligned} \tag{64}$$

Using the same type of reasoning as for the upper bound, one can show that equation (64) yields:

$$\begin{aligned} y_i \Xi_{i,j,r}^{(t+1)} &\geq -\tilde{O}(\sigma\sigma_0\sqrt{d}) - \tilde{O}\left(\frac{\sigma_0\sigma^4 d^2\beta^3}{\alpha}\right) \\ &- \tilde{O}\left(\frac{P\sigma_0^2\sigma^4 d^{3/2}}{\alpha}\right) + \tilde{O}\left(\frac{P\sigma_0\sigma^4 d^{3/2}\beta^3}{\alpha}\right) \\ &- \tilde{O}(P\sigma^2\sqrt{d}) \\ &\geq -\tilde{O}(\sigma\sigma_0\sqrt{d}). \end{aligned} \tag{65}$$

equation (65) shows the induction hypothesis for $t+1$.

□

Proof of Induction hypothesis C.2. In this section, we prove the induction hypotheses for the signal $c^{(t)}$.

Proof of $c^{(t+1)} \geq -\tilde{O}(\sigma_0)$. We know that with high probability, $c^{(0)} \geq -\tilde{O}(\sigma_0)$. By Lemma E.1, $c^{(t)}$ is a non-decreasing sequence and therefore, we always have $c^{(t)} \geq -\tilde{O}(\sigma_0)$.

Proof of $c^{(t+1)} \leq \tilde{O}(1/\alpha)$. Using the same proof as the one for Lemma 4.6, we prove the induction hypothesis for $t+1$.

□

Proof of Induction hypothesis C.3. $\alpha \min\{1, (c^{(t)})^2\alpha^2\} \geq (\Xi_{i,j,r}^{(t)})^2$ is true for all t. Indeed, we proved in Lemma 4.1 that after T_0 iterations, $c^{(t)} \geq \tilde{\Omega}(1/\alpha)$. Moreover, we proved Induction hypothesis C.1 claiming that $|\Xi_{i,j,r}^{(t)}| \leq \tilde{O}(1)$ Therefore, we have $\alpha \min\{1, (c^{(t)})^2\alpha^2\} \geq (\Xi_{i,j,r}^{(t)})^2$.

□

F LEARNING WITH GD+M

In this section, we prove the Lemmas in Section 5 and Theorem 3.2.

F.1 LEARNING SIGNAL WITH GD+M

To track the amount of signal learnt by GD, we make use of the following update.

Lemma F.1 (Signal momentum). *For all $t \geq 0$ and $r \in [m]$, the signal momentum in equation (GDM-S) is equal to:*

$$\mathcal{G}_r^{(t+1)} = \gamma \mathcal{G}_r^{(t)} - 3(1-\gamma)\left(\alpha^3 \nu_1^{(t)} + \beta^3 \nu_2^{(t)}\right)(c_r^{(t)})^2.$$

We can further simplify this update as:

$$\mathcal{G}_r^{(t+1)} = \gamma \mathcal{G}_r^{(t)} - \Theta(1-\gamma)\left(\alpha^3(1-\hat{\mu})\widehat{\ell}^{(t)}(\alpha) + \beta^3 \hat{\mu}\widehat{\ell}^{(t)}(\beta)\right)(c_r^{(t)})^2.$$

Proof of Lemma F.1. By definition of the momentum update, we have: $g_r^{(t+1)} = \gamma g_r^{(t)} + (1-\gamma)\nabla_{w_r}\widehat{L}(W^{(t)})$. We project this update onto w^* and use Lemma D.2 to get:

$$\mathcal{G}_r^{(t+1)} = \gamma \mathcal{G}_r^{(t)} - 3(1-\gamma)\left(\alpha^3 \nu_1^{(t)} + \beta^3 \nu_2^{(t)}\right)(c_r^{(t)})^2. \tag{66}$$

By applying Lemma I.1, we have $\nu_1^{(t)} = \Theta(1-\hat{\mu})\widehat{\ell}^{(t)}(\alpha)$ and $\nu_2^{(t)} = \Theta(\hat{\mu})\widehat{\ell}^{(t)}(\beta)$. Plugging this observation in equation (66) yields the desired result. □

F.1.1 EARLY STAGES OF THE LEARNING PROCESS $t \in [0, \mathcal{T}_0]$: LEARNING $\mathcal{Z}_1$ DATA

Similarly to GD, since we initialize $w_r^{(0)} \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I})$ with σ_0 small, the sigmoid terms $\widehat{\ell}^{(t)}(\alpha)$ and $\widehat{\ell}^{(t)}(\beta)$ in the momentum are large at early iterations. As $c_r^{(t)}$ is non-decreasing (by Lemma I.2), $\widehat{\ell}^{(t)}(\alpha)$ eventually becomes small at a time $\mathcal{T}_0 > 0$. We therefore simplify the signal momentum update for $t \in [0, \mathcal{T}_0]$.

Lemma F.2 (Signal momentum at early iterations). *Let $\mathcal{T}_0 > 0$ the time where there exists $s \in [m]$ such that $c_s^{(t)} \geq \tilde{\Omega}(1/\alpha)$. Then, for $t \in [0, \mathcal{T}_0]$ and $r \in [m]$, the signal momentum is simplified as:*

$$\mathcal{G}_r^{(t+1)} = \gamma \mathcal{G}_r^{(t)} - \Theta(\alpha^3(1-\gamma))(c_r^{(t)})^2. \tag{67}$$

Proof of Lemma F.2. From Lemma F.1, we can simplify the momentum update as:

$$-\hat{\mu}\widehat{\ell}^{(t)}(\beta)(c_r^{(t)})^2 \leq \mathcal{G}_r^{(t+1)} - \gamma \mathcal{G}_r^{(t)} + \Theta(1-\gamma)\alpha^3(1-\hat{\mu})\widehat{\ell}^{(t)}(\alpha)(c_r^{(t)})^2 \leq 0. \tag{68}$$

For $t \in [0, \mathcal{T}_0]$, we know that for all $s \in [m]$, we have $c_s^{(t)} \leq \frac{\tilde{O}(1)}{m^{1/3}\alpha} = \frac{\tilde{O}(1)}{\alpha}$ (since $m = \tilde{O}(1)$). Thus, we have:

$$\frac{1}{1+\exp(\tilde{\Omega}(1))} \leq \widehat{\ell}^{(t)}(\alpha) = \frac{1}{1+\exp\left(\sum_{s=1}^m \alpha^3 (c_s^{(t)})^3\right)} \leq 1. \tag{69}$$

By Remark 1, we know that the sigmoid is small only when we have $\frac{1}{1+\exp(\tilde{\Omega}(1))}$. From equation (69), we have

$$\widehat{\ell}^{(t)}(\alpha) = \Theta(1). \tag{70}$$

Besides, we have:

$$\frac{1}{1+\exp(\tilde{\Omega}(1))} \leq \frac{1}{1+\exp(\tilde{\Omega}(\beta^3/\alpha^3))} \leq \widehat{\ell}^{(t)}(\beta) = \frac{1}{1+\exp\left(\sum_{s=1}^m \beta^3 (c_s^{(t)})^3\right)} \leq 1. \tag{71}$$

From equation (71), we have:

$$\widehat{\ell}^{(t)}(\beta) = \Theta(1). \tag{72}$$

Plugging equation (70) and equation (72) in equation (68) yields the desired result. □

We now prove Lemma 5.1 that quantifies the amount of signal learnt by GD when the gradient is large.

Lemma 5.1. *For all $r \in [m]$ and $t \geq 0$, as long as $c^{(t)} \leq \tilde{O}(1/\alpha)$, the momentum update equation (GDM-S) is simplified as:*

$$-\mathcal{G}_r^{(t+1)} = -\gamma \mathcal{G}_r^{(t)} + (1-\gamma)\Theta(\alpha^3)(c_r^{(t)})^2$$

Consequently, after $\mathcal{T}_0 = \tilde{\Theta}\left(\frac{1}{\sigma_0 \alpha^2} + \frac{1}{1-\gamma}\right)$ iterations, for all $t \in [\mathcal{T}_0, T]$, we have $c^{(t)} \geq \tilde{\Omega}(1/\alpha)$.

Proof of Lemma 5.1. By Lemma F.2, the signal update for $t \in [0, \mathcal{T}_0]$ satisfies:

$$\begin{cases} \mathcal{G}_r^{(t+1)} & = \gamma \mathcal{G}_r^{(t)} - (1-\gamma)\tilde{\Theta}(\alpha^3)(c_r^{(t)})^2 \\ c_r^{(t+1)} & = c_r^{(t)} - \eta \mathcal{G}_r^{(t+1)} \end{cases}. \tag{73}$$

As $c_r^{(t)}$ is non-decreasing (by Lemma I.2), it will eventually reach $\tilde{\Omega}(1/\alpha)$. We can use the arguments as in the proof of Lemma 4.1 to argue that there exists an index r such that $c_r^{(t)} > 0$. Among all the possible indices, we focus on $r = r_{\max}$, where $r_{\max} = \operatorname{argmax}_{r \in [m]} c_r^{(0)}$.

To find $\mathcal{T}_0$, we apply the Tensor Power Method (Lemma J.17) to equation (73). Setting $\upsilon = \tilde{O}(1/\alpha)$ in Lemma J.17, we deduce that the time t_0 is for $\delta \in (0,1)$

$$t_0 = \frac{1}{1-\gamma}\left\lceil \frac{\log(\tilde{O}(1/\alpha))}{\log(1+\delta)} \right\rceil + \frac{1+\delta}{\eta(1-e^{-1})\alpha^3 c^{(0)}},$$

Since $c^{(0)} = \tilde{\Theta}(\sigma_0)$ with high probability, we can set up $\mathcal{T}_0$ such that $t_0 \leq \mathcal{T}_0 = \tilde{\Theta}\left(\frac{1}{1-\gamma} + \frac{1}{\alpha^3 \sigma_0}\right)$. □

F.1.2 Late stages of learning process $t \in [\mathcal{T}_0, T]$: learning $\mathcal{Z}_2$ data

We now show that contrary to GD, GD+M still has a large momentum in the w^* direction. In other words, we want to show that $-\mathcal{G}^{(t)}$ is still large for $t \in [\mathcal{T}, T]$. Given that the small margin and large margin data share the same feature w^*, this large momentum helps to learn $\mathcal{Z}_2$.

Before proving such result, we need some intermediate lemmas.

Lemma F.3. *Let $\mathcal{T} > 0$ such that $-\mathcal{G}^{(\mathcal{T})} \leq \tilde{O}(\sqrt{1-\gamma}/\alpha)$. Then, for all $t' \leq \mathcal{T}$, we have:*

$$-\mathcal{G}^{(t')} \leq \frac{\tilde{O}(\sqrt{1-\gamma})}{\alpha \gamma^{\mathcal{T}-t'}}.$$

Proof of Lemma F.3. Using the momentum update rule, we know that:

$$-\mathcal{G}^{(\mathcal{T})} = -\gamma^{\mathcal{T}-t'}\mathcal{G}^{(t')} - (1-\gamma)\sum_{\tau=t'}^{\mathcal{T}-1} \gamma^{\mathcal{T}-\tau}\mathscr{G}^{(\tau)}. \tag{74}$$

Since $-\mathscr{G}^{(\tau)} > 0$ for all $\tau \geq 0$, equation (74) implies $-\gamma^{\mathcal{T}-t'}\mathcal{G}^{(t')} \leq -\mathcal{G}^{(\mathcal{T})}$. Using $-\mathcal{G}^{(\mathcal{T})} \leq \tilde{O}(\sqrt{1-\gamma}/\alpha)$, we obtain the aimed result. □

Lemma F.4. *Let $\mathcal{T}_0$ be the first iteration where $c^{(t)} > \tilde{\Omega}(1/\alpha)$. Assume that $\mathcal{G}^{(\mathcal{T}_0)} \leq \tilde{O}(\sqrt{1-\gamma}/\alpha)$ Then, we have:*

$$c^{(t)} \geq \tilde{\Omega}(1/\alpha).$$

for all $t \in \left[\mathcal{T}_0 - \frac{1}{\sqrt{1-\gamma}}, \mathcal{T}_0\right]$

Proof of Lemma F.4. Let's define $t' := \mathcal{T}_0 - \frac{1}{\sqrt{1-\gamma}}$. We start by summing the GD+M update equation (GDM-S) for $\tau = t', \ldots, \mathcal{T}_0$ to get

$$c^{(\mathcal{T}_0)} = c^{(t')} - \eta \sum_{\tau=t'}^{\mathcal{T}_0-1} \mathcal{G}^{(\tau)}. \tag{75}$$

Applying Lemma F.3 to bound the momentum gradient, we further bound equation (75) to get:

$$\begin{aligned} c^{(t')} &= c^{(\mathcal{T}_0)} - \eta \sum_{\tau=t'}^{\mathcal{T}_0-1} \mathcal{G}^{(\tau)} \\ &\geq c^{(\mathcal{T}_0)} - \eta\tilde{O}(\sqrt{1-\gamma}/\alpha) \sum_{\tau=t'}^{\mathcal{T}_0-1} \frac{1}{\gamma^{\mathcal{T}_0-\tau}} \\ &= c^{(\mathcal{T}_0)} - \eta\tilde{O}(\sqrt{1-\gamma}/\alpha) \sum_{j=1}^{\mathcal{T}_0-t'} \gamma^{-j} \\ &= c^{(\mathcal{T}_0)} - \eta\tilde{O}(\sqrt{1-\gamma}/\alpha) \frac{1-\gamma^{\mathcal{T}_0-t'}}{1-\gamma}. \end{aligned} \tag{76}$$

We now use the fact that $\mathcal{T}_0 - t' = \frac{1}{\sqrt{1-\gamma}}$ in equation (76) to get:

$$c^{(t')} \geq c^{(\mathcal{T}_0)} - \eta\tilde{O}(1) \frac{1-\gamma^{\frac{1}{\sqrt{1-\gamma}}}}{\sqrt{1-\gamma}/\alpha}. \tag{77}$$

Since $\gamma = 1 - \varepsilon$ with $\varepsilon \ll 1$, we linearize the right-hand side in equation (77) to obtain:

$$\begin{aligned} c^{(t')} &\geq c^{(\mathcal{T}_0)} - \tilde{O}(\eta) \frac{1-(1-\varepsilon)^{\frac{1}{\sqrt{\varepsilon}}}}{\sqrt{\epsilon}\alpha} \\ &= c^{(\mathcal{T}_0)} - \eta\tilde{O}(\eta) \frac{1-(1-\varepsilon)^{\frac{1}{\sqrt{\varepsilon}}}}{\sqrt{\epsilon}\alpha} \\ &= c^{(\mathcal{T}_0)} - \tilde{O}(\eta/\alpha). \end{aligned} \tag{78}$$

Since $c^{(\mathcal{T}_0)} \geq \tilde{\Omega}(1/\alpha)$ and $\tilde{O}(\eta) \leq 0.5\tilde{\Omega}(1)$, equation (78) finally yields the desired result. □

Using Lemma F.4, we can therefore show that once we learn $\mathcal{Z}_1$, the signal momentum still stays large.

Lemma 5.2. *Let* $\mathcal{T}_0 = \tilde{\Theta}\left(\frac{1}{\sigma_0\alpha^3} + \frac{1}{1-\gamma}\right)$. *For* $t \in [\mathcal{T}_0, T]$, *we have* $\mathcal{G}^{(t)} \geq \tilde{\Omega}(\sqrt{1-\gamma}/\alpha)$.

Proof of Lemma 5.2. By contradiction, let's assume that $-\mathcal{G}^{(\mathcal{T}_0)} \leq \tilde{O}(\sqrt{1-\gamma}/\alpha)$. Let's define $t' := \mathcal{T}_0 - \frac{1}{\sqrt{1-\gamma}}$. Since $-\mathcal{G}^{(t')} \geq 0$, $-\mathcal{G}^{(\mathcal{T}_0)}$ is bounded as:

$$-\mathcal{G}^{(\mathcal{T}_0)} \geq \tilde{\Theta}(1)(1-\gamma)\alpha^3 \sum_{\tau=t'}^{\mathcal{T}_0-1} \gamma^{\mathcal{T}_0-1-\tau}(c_r^{(\tau)})^2. \tag{79}$$

Using Lemma F.4, we bound $(c_r^{(\tau)})$ in equation (79) and get:

$$\begin{aligned} -\mathcal{G}^{(\mathcal{T}_0)} &\geq (1-\gamma)\tilde{\Omega}(\alpha) \sum_{\tau=t'}^{\mathcal{T}_0-1} \gamma^{\mathcal{T}_0-1-\tau} \\ &= (1-\gamma)\tilde{\Omega}(\alpha) \sum_{\tau=0}^{\mathcal{T}_0-1-t'} \gamma^{j} \\ &= \tilde{\Omega}(\alpha)(1-\gamma^{\mathcal{T}_0-t'}). \\ &= \tilde{\Omega}(\alpha)(1-\gamma^{1/\sqrt{1-\gamma}}) \end{aligned} \tag{80}$$

Since $\gamma = 1 - \varepsilon$ with $\epsilon \ll 1$, we have $(1 - \gamma^{1/\sqrt{1-\gamma}}) \geq \sqrt{1-\gamma}$. Therefore, we proved that $\mathcal{G}^{(\mathcal{T}_0)} \geq \tilde{O}(\sqrt{1-\gamma})$ which is a contradiction. □

Since the signal momentum is large (Lemma 5.2), we want to argue that GD+M keeps learning the feature to eventually have a large signal.

Lemma 5.3. *Let $\mathcal{T}_0 = \tilde{\Theta}\left(\frac{1}{\sigma_0 \alpha^3} + \frac{1}{1-\gamma}\right)$. After $\mathcal{T}_1 = \mathcal{T}_0 + \tilde{\Theta}\left(\frac{1}{1-\gamma}\right)$ iterations, for $t \in [\mathcal{T}_1, T]$, we have $c^{(t)} \geq \tilde{\Omega}\left(\frac{1}{\sqrt{1-\gamma}\alpha}\right)$. Our choice of parameter in Section 2, this implies $c^{(t)} \geq \tilde{\Omega}(1/\beta)$.*

Proof of Lemma 5.3. Let $\mathcal{T}_1 \in [T]$ such that $\mathcal{T}_0 < \mathcal{T}_1$. From the signal momentum update, we deduce:

$$-\mathcal{G}^{(\mathcal{T}_1)} \geq -\gamma^{\mathcal{T}_1 - \mathcal{T}_0}\mathcal{G}^{(\mathcal{T}_0)} + \sum_{\tau=\mathcal{T}_0}^{\mathcal{T}_1} \gamma^{\mathcal{T}_1 - \tau}(\mathscr{G}^{(\tau)})^2 \geq -\gamma^{\mathcal{T}_1 - \mathcal{T}_0}\mathcal{G}^{(\mathcal{T}_0)}. \tag{81}$$

We now apply Lemma 5.2 to bound $-\mathcal{G}^{(\mathcal{T}_1)}$ in equation (81) and get:

$$-\mathcal{G}^{(\mathcal{T}_1)} \geq \gamma^{\mathcal{T}_1 - \mathcal{T}_0}\tilde{\Omega}(\sqrt{1-\gamma}/\alpha). \tag{82}$$

We would like to find the time such that $\gamma^{\mathcal{T}_1 - \mathcal{T}_0}$ is a constant factor $a \leq 1$ i.e. such that

$$\gamma^{\mathcal{T}_1 - \mathcal{T}_0} = a \iff \mathcal{T}_1 - \mathcal{T}_0 = \frac{-\log(a)}{-\log(\gamma)} \leq \frac{\log(a)}{1-\gamma}, \tag{83}$$

where we used the fact that $\log(x) \leq x - 1$ for $x > 0$ in the last inequality. Therefore, we proved that $\mathcal{T}_1 = \mathcal{T}_0 + \tilde{O}(\frac{1}{1-\gamma})$ and

$$-\mathcal{G}^{(\mathcal{T}_1)} \geq -a\mathcal{G}^{(\mathcal{T}_0)} = \tilde{\Omega}\left(\frac{1}{\sqrt{1-\gamma}\alpha}\right). \tag{84}$$

Let $t \in (\mathcal{T}_1, T]$. Using equation (GDM-S) update rule, we have $c^{(t)} = c^{(t-1)} - \eta\mathcal{G}^{(t)}$. Since $t \geq \mathcal{T}_0$, we have $c^{(t-1)} \geq \tilde{\Omega}(1/\alpha) > 0$ which implies

$$c^{(t)} \geq -\eta\mathcal{G}^{(t)}. \tag{85}$$

Moreover, since $-\mathcal{G}^{(t)}$ is a non-decreasing sequence, we have $-\mathcal{G}^{(t)} \geq -\mathcal{G}^{(\mathcal{T}_1)}$. Plugging this bound in equation (85) yields

$$c^{(t)} \geq -\eta\mathcal{G}^{(\mathcal{T}_1)}. \tag{86}$$

Using $\eta = \tilde{\Theta}(1)$ and plugging equation (84) in equation (86) lead to the final result.

□

Lemma 5.3 implies that after $\mathcal{T}_1$ iterations, the learnt signal is very large which implies that the full derivative quickly decreases. This statement is formally made in Lemma 5.4. Before proving this result, we need an auxiliary lemma that connects the signal momentum and the full derivative $\nu^{(t)}$.

Lemma F.5 (Bound on signal momentum)**.** *For $t \in [\mathcal{T}_1, T]$, the signal momentum is bounded as*

$$-\mathcal{G}^{(t+1)} \geq -\gamma\mathcal{G}^{(t)} + (1-\gamma)\Omega\left(\nu^{(t)}\beta\right)$$

Proof of Lemma F.5. From Lemma F.1 we know that the signal momentum is equal to

$$-\mathcal{G}^{(t+1)} = -\gamma\mathcal{G}^{(t)} + 3(1-\gamma)\left(\alpha^3\nu_1^{(t)} + \beta^3\nu_2^{(t)}\right)(c^{(t)})^2. \tag{87}$$

Since $\beta \leq \alpha$, equation (87) becomes

$$-\mathcal{G}^{(t+1)} \geq -\gamma\mathcal{G}^{(t)} + \Theta(1)(1-\gamma)\beta^3\nu^{(t)}(c^{(t)})^2. \tag{88}$$

We finally apply Lemma 5.3 to bound $c^{(t)}$ in equation (88) to obtain the desired result. □

We now present the proof of Lemma 5.4.

Lemma 5.4. *Let* $\mathcal{T}_0 = \tilde{\Theta}\left(\frac{1}{\eta\sigma_0\alpha^3} + \frac{1}{1-\gamma}\right)$. *After* $\mathcal{T}_1 = \mathcal{T}_0 + \tilde{\Theta}\left(\frac{1}{1-\gamma}\right)$ *iterations, for* $t \in [\mathcal{T}_1, T]$, $\nu^{(t)} \leq \tilde{O}\left(\frac{1}{\eta(t-\mathcal{T}_1+1)\beta}\right)$.

Proof of Lemma 5.4 . Lemma F.5 provides an upper bound on $\nu^{(t)}$ since:

$$\nu^{(t)} \leq \tilde{O}\left(\frac{1}{(1-\gamma)\beta}\right)(\mathcal{G}^{(t+1)} - \gamma\mathcal{G}^{(t)}). \tag{89}$$

We now would like to give a convergence rate on the iterates $\mathcal{G}^{(t+1)} - \gamma\mathcal{G}^{(t)}$. Since Lemma I.5 gives a rate on the loss function, we now need connect the momentum increment to a loss term. Applying Lemma F.1, we have:

$$\begin{aligned}\mathcal{G}^{(t+1)} - \gamma\mathcal{G}^{(t)} &\leq \sum_{r=1}^{m} |\mathcal{G}_r^{(t+1)} - \gamma\mathcal{G}_r^{(t)}| \\ &= \tilde{\Theta}(1)(1-\gamma)\sum_{r=1}^{m}\left((1-\hat{\mu})\alpha^3\widehat{\ell}^{(t)}(\alpha) + \hat{\mu}\beta^3\widehat{\ell}^{(t)}(\beta)\right)(c_r^{(t)})^2. \end{aligned} \tag{90}$$

We want to now show that for $t \in [\mathcal{T}_1, T]$, we have:

$$(1-\hat{\mu})\alpha^3\widehat{\ell}^{(t)}(\alpha) \leq \hat{\mu}\beta^3\widehat{\ell}^{(t)}(\beta). \tag{91}$$

Indeed, by Lemma 5.3, we have:

$$(1-\hat{\mu})\alpha^3\widehat{\ell}^{(t)}(\alpha) \leq \frac{\Theta(\alpha^3)}{1+\exp(\tilde{\Omega}(\alpha^3/\beta^3))}, \tag{92}$$

$$\hat{\mu}\beta^3\widehat{\ell}^{(t)}(\beta) \geq \frac{\hat{\mu}\beta^3}{1+\exp(\tilde{O}(1))}. \tag{93}$$

Thus, combining equation (92) and equation (93) yields:

$$\frac{(1-\hat{\mu})\alpha^3\widehat{\ell}^{(t)}(\alpha)}{\hat{\mu}\beta^3\widehat{\ell}^{(t)}(\beta)} \leq \frac{\Theta(\alpha^3)}{\hat{\mu}\beta^3}\frac{1+\exp(\tilde{O}(1))}{1+\exp(\tilde{\Omega}(\alpha^3/\beta^3))}. \tag{94}$$

Since $\alpha = d^{0.49}$, $\beta = \frac{\alpha}{\text{polylog(d)}\sqrt{d}}$ and $\hat{\mu} = 1/\text{poly}(d)$, we finally bound equation (94) as:

$$\frac{(1-\hat{\mu})\alpha^3\widehat{\ell}^{(t)}(\alpha)}{\hat{\mu}\beta^3\widehat{\ell}^{(t)}(\beta)} \leq \frac{\text{poly}(d)\exp(O(1)\text{polylog}(d))}{\exp(\Omega(1)\text{polylog}(d)d^{3/2})} \leq 1. \tag{95}$$

Therefore, plugging equation (91) in equation (90) yields:

$$\mathcal{G}^{(t+1)} - \gamma\mathcal{G}^{(t)} \leq 2\tilde{\Theta}(1)(1-\gamma)\sum_{r=1}^{m}\hat{\mu}\beta^3\widehat{\ell}^{(t)}(\beta)(c_r^{(t)})^2. \tag{96}$$

We now apply Lemma J.20 to link equation (96) with a loss term. By Lemma 5.3, we have:

$$\tilde{\Omega}(1) \leq \tilde{\Omega}(1) - m\tilde{O}(\sigma_0) \leq \sum_{r=1}^{m}\beta c_r^{(t)} \leq \tilde{O}(m) \leq \tilde{O}(1). \tag{97}$$

Therefore, applying Lemma J.20 in equation (96) gives:

$$\mathcal{G}^{(t+1)} - \gamma\mathcal{G}^{(t)} \leq 40\hat{\mu}\tilde{\Theta}(1)(1-\gamma)\frac{m\beta e^{m\tilde{O}(\sigma_0)}}{\tilde{\Omega}(1)}\widehat{\mathcal{L}}^{(t)}(\beta) \leq \tilde{O}(\beta)\hat{\mu}(1-\gamma)\widehat{\mathcal{L}}^{(t)}(\beta). \tag{98}$$

Thus, plugging equation (98) in equation (89) yields:

$$\nu^{(t)} \leq \tilde{O}(1)\hat{\mu}\widehat{\mathcal{L}}^{(t)}(\beta). \tag{99}$$

We finally apply Lemma I.5 to bound the loss term in equation (99) and get the desired result. □

After $\mathcal{T}_1$ iterations, the gradient is now very small and GD+M cannot update its weights. At this time, the noise component learnt by GD+M stays very small.

Lemma 5.5. *Let $i \in [N]$, $j \in [P] \backslash \{P(X_i)\}$ and $r \in [m]$. For $t \geq 0$, equation (GDM-N) can be rewritten as $|G_{i,j,r}^{(t+1)}| \leq \gamma |G_{i,j,r}^{(t)}| + (1-\gamma)\tilde{O}(\sigma_0^2 \sigma^4 d^2)\nu^{(t)}$. As a consequence, after $t \in [\mathcal{T}_1, T]$ iterations, we thus have $|\Xi_{i,j,r}^{(t)}| \leq \tilde{O}(\sigma_0 \sigma \sqrt{d})$.*

Proof of Lemma 5.5 . This Lemma is intended to prove Induction hypothesis C.4. At time $t = 0$, we have $|\Xi_{i,j,r}^{(0)}| \leq \tilde{O}(\sigma \sigma_0 \sqrt{d})$ by Lemma J.7. Assume that Induction hypothesis C.4 is true for $t \in [\mathcal{T}_1, T]$. Now, let's prove this induction hypothesis for time $t+1$. Let We remind that equation (GDM-N) update rule is

$$\Xi_{i,j,r}^{(t+1)} = \Xi_{i,j,r}^{(t)} - \eta G_{i,j,r}^{(t+1)}. \tag{100}$$

We now use Induction hypothesis C.4 to bound $\Xi_{i,j,r}^{(t)}$ in equation (100):

$$|\Xi_{i,j,r}^{(t+1)}| \leq \tilde{O}(\sigma \sigma_0 \sqrt{d}) + \eta |G_{i,j,r}^{(t)}|. \tag{101}$$

To bound $G_{i,j,r}^{(t+1)}$, we use Lemma I.4. We have:

$$|G_{i,j,r}^{(t)}| \leq (1-\gamma)\tilde{O}(\sigma^4 \sigma_0^2 d^2) \left(\sum_{\tau=0}^{\mathcal{T}_1 - 1} \gamma^{t-1-\tau} \nu^{(\tau)} + \sum_{\tau=\mathcal{T}_1}^{t-1} \gamma^{t-1-\tau} \nu^{(\tau)} \right). \tag{102}$$

On one hand, we can bound the first sum in equation (102) as:

$$\sum_{\tau=0}^{\mathcal{T}_1 - 1} \gamma^{t-1-\tau} \nu^{(\tau)} \leq \sum_{\tau=0}^{\mathcal{T}_1 - 1} \gamma^{t-1-\tau} = \gamma^{t-\mathcal{T}_1} \frac{1-\gamma^{\mathcal{T}_1}}{1-\gamma} \leq \frac{\gamma^{t-\mathcal{T}_1}}{1-\gamma}. \tag{103}$$

On the other hand, using Lemma 5.4, the second sum in equation (102) is bounded as:

$$\sum_{\tau=\mathcal{T}_1}^{t-1} \gamma^{t-1-\tau} \nu^{(\tau)} \leq \frac{1}{\beta} \sum_{\tau=\mathcal{T}_1}^{t-1} \frac{\gamma^{t-1-\tau}}{\tau - \mathcal{T}_1 - 1} \leq \frac{1}{\beta} \sum_{\tau=1}^{t-\mathcal{T}_1} \gamma^{t-\tau} \leq \frac{1}{\beta} \frac{1}{1-\gamma}. \tag{104}$$

Plugging equation (103) and equation (104) in equation (102) yields

$$|G_{i,j,r}^{(t)}| \leq \tilde{O}(\sigma^4 \sigma_0^2 d^2) \left(\gamma^{t-\mathcal{T}_1} + \frac{1}{\beta} \right) \leq \frac{\tilde{O}(\sigma^4 \sigma_0^2 d^2)}{\beta} \leq \tilde{O}(\sigma \sigma_0 \sqrt{d}), \tag{105}$$

where we used $\beta = d^{-0.01}/\mathrm{polylog}(d) \leq \sigma \sigma_0 \sqrt{d}$ in the last inequality. Plugging equation (105) in equation (101) shows that Induction hypothesis C.4 is true for $t+1$.

□

We proved that the weights learnt by GD+M satisfy for $r \in [m]$

$$w_r^{(T)} = c_r^{(T)} w^* + v_r^{(T)}, \tag{106}$$

where one of the $c_r^{(T)} = \tilde{\Omega}(1/\beta)$ (Lemma 5.3) and $v_r^{(T)} \in \mathrm{span}(X_i[j]) \subset \mathrm{span}(w^*)^{\perp}$. By Lemma 5.5, since $\Xi_{i,j,r}^{(t)} \leq \tilde{O}(\sigma_0)$, we have $\|v_r^{(T)}\|_2 \leq 1$. We are now ready to prove the generalization achieved by GD+M and stated in Theorem 3.2.

Theorem 3.2. *Assume that we run GD+M on equation (P) for T iterations with parameters set as in Parametrization 2.1. With probability at least $1 - o(1)$, the weights learned by GD+M*

1. *(at least for one of them) is highly correlated with the feature: $c^{(T)} > \tilde{\Omega}(1/\beta)$.*

2. *are barely correlated with noise: for all $r \in [m]$, for all $i \in [N]$ and $j \in [P]$. $|\Xi_{i,j,r}^{(T)}| \leq \tilde{O}(\sigma_0)$.*

*Consequently, the training loss and the test error are **at most** $O(\mu/\mathrm{poly}(d))$.*

Proof of Theorem 3.2. We now bound the training and test error achieved by GD+M at time T.

Train error. Lemma I.5 provides a convergence bound on the fake loss. Indeed, we know that:

$$(1-\hat{\mu})\widehat{\mathcal{L}}^{(T)}(\alpha)+\hat{\mu}\widehat{\mathcal{L}}^{(T)}(\beta)\leq\tilde{O}\left(\frac{1}{\eta(T-\mathcal{T}_1+1)}\right). \tag{107}$$

Using Lemma J.24 along with Induction hypothesis C.4, we can actually lower bound the loss term in equation (107) by the true loss.

$$\Theta(1)\widehat{L}(W^{(T)})\leq(1-\hat{\mu})\widehat{\mathcal{L}}^{(T)}(\alpha)+\hat{\mu}\widehat{\mathcal{L}}^{(T)}(\beta). \tag{108}$$

Combining equation (107) and equation (108), we obtain a bound on the training loss.

$$\widehat{L}(W^{(T)})\leq\frac{\tilde{\Theta}(1)}{\eta(T-\mathcal{T}_1+1)}. \tag{109}$$

Plugging $T\geq\text{poly}(d)N/\eta$ and $\mu=\Theta(1/N)$ in equation (109) yields:

$$\widehat{L}(W^{(T)})\leq\tilde{O}\left(\frac{1}{\text{poly}(d)N}\right)\leq\tilde{O}\left(\frac{\mu}{\text{poly}(d)}\right). \tag{110}$$

Test error. Let (X,y) be a datapoint. We remind that $X=(X[1],\dots,X[P])$ where $X[P(X)]=\theta y w^*$ and $X[j]\sim\mathcal{N}(0,\sigma^2\mathbf{I})$ for $j\in[P]\backslash\{P(X)\}$. We bound the test error as follows:

$$\begin{aligned}\mathscr{L}(f_{W^T})&=\mathbb{E}_{\substack{\mathcal{Z}\sim\mathcal{D}\\(X,y)\sim\mathcal{Z}}}[\mathbf{1}_{yf_{W^T}(X)<0}]\\&=\mathbb{E}_{(X,y)\sim\mathcal{Z}_1}[\mathbf{1}_{yf_{W^T}(X)<0}]\mathbb{P}[\mathcal{Z}_1]+\mathbb{E}_{(X,y)\sim\mathcal{Z}_2}[\mathbf{1}_{yf_{W^T}(X)<0}]\mathbb{P}[\mathcal{Z}_2]\\&=(1-\hat{\mu})\mathbb{P}[yf_{W^T}((X)<0|(X,y)\sim\mathcal{Z}_1]+\hat{\mu}\mathbb{P}[yf_{W^T}(X)<0|(X,y)\sim\mathcal{Z}_2].\end{aligned} \tag{111}$$

We now want to compute the probability terms in equation (111). We remind that $yf_{W^T}(X)$ is given by

$$\begin{aligned}yf_{W^T}(X)&=y\sum_{s=1}^{m}\sum_{j=1}^{P}\langle w_s^{(T)},X[j]\rangle^3\\&=\theta^3\sum_{s=1}^{m}(c_s^{(T)})^3+y\sum_{s=1}^{m}\sum_{j\neq P(X)}\langle v_s^{(T)},X[j]\rangle^3\\&\geq\theta^3(c^{(T)})^3+y\sum_{s=1}^{m}\sum_{j\neq P(X)}\langle v_s^{(T)},X[j]\rangle^3.\end{aligned} \tag{112}$$

We now apply Lemma 5.3, equation (112) is finally bounded as:

$$yf(X)\geq\Omega\left(\frac{\theta^3}{\beta^3}\right)+y\sum_{s=1}^{m}\sum_{j\neq P(X)}\langle v_s^{(T)},X[j]\rangle^3. \tag{113}$$

Therefore, using equation (113), we upper bound the test error equation (111) as:

$$\begin{aligned}\mathscr{L}(f_{W^T})\leq(1-\hat{\mu})\mathbb{P}&\left[y\sum_{s=1}^{m}\sum_{j\neq P(X)}\langle v_s^{(T)},X[j]\rangle^3\leq-\tilde{\Omega}\left(\frac{\alpha^3}{\beta^3}\right)\right]\\&+\hat{\mu}\mathbb{P}\left[y\sum_{s=1}^{m}\sum_{j\neq P(X)}\langle v_s^{(T)},X[j]\rangle^3\leq-\tilde{\Omega}(1)\right].\end{aligned} \tag{114}$$

Since y is taken uniformly from $\{-1,1\}$, we further simplify equation (114) as:

$$\begin{aligned}\mathscr{L}(f_{W^T})\leq&\frac{1-\hat{\mu}}{2}\left(\mathbb{P}\left[\sum_{s=1}^{m}\sum_{j\neq P(X)}\langle v_s^{(T)},X[j]\rangle^3\leq-\tilde{\Omega}\left(\frac{\alpha^3}{\beta^3}\right)\right]+\mathbb{P}\left[\sum_{s=1}^{m}\sum_{j\neq P(X)}\langle v_s^{(T)},X[j]\rangle^3\geq\tilde{\Omega}\left(\frac{\alpha^3}{\beta^3}\right)\right]\right)\\&+\frac{\hat{\mu}}{2}\left(\mathbb{P}\left[\sum_{s=1}^{m}\sum_{j\neq P(X)}\langle v_s^{(T)},X[j]\rangle^3\leq-\tilde{\Omega}(1)\right]+\mathbb{P}\left[\sum_{s=1}^{m}\sum_{j\neq P(X)}\langle v_s^{(T)},X[j]\rangle^3\geq\tilde{\Omega}(1)\right]\right).\end{aligned} \tag{115}$$

We know that $\langle v_s^{(T)}, X[j]\rangle \sim \mathcal{N}(0, \|v_s^{(T)}\|_2^2\sigma^2)$. Therefore, $\langle v_s^{(T)}, X[j]\rangle^3$ is the cube of a centered Gaussian.This random variable is symmetric. By Lemma J.1, we know that $\sum_{s=1}^m \sum_{j\neq P(X)} \langle v_s^{(T)}, X[j]\rangle^3$ is also symmetric. Therefore, we simplify equation (115) as:

$$\begin{aligned}\mathscr{L}(f_{W^T}) \leq (1-\hat{\mu})\mathbb{P}\left[\sum_{s=1}^m \sum_{j\neq P(X)} \langle v_s^{(T)}, X[j]\rangle^3 \geq \tilde{\Omega}\left(\frac{\alpha^3}{\beta^3}\right)\right] \\ + \hat{\mu}\mathbb{P}\left[\sum_{s=1}^m \sum_{j\neq P(X)} \langle v_s^{(T)}, X[j]\rangle^3 \geq \tilde{\Omega}(1)\right].\end{aligned} \tag{116}$$

From Lemma J.14, we know that $\sum_{s=1}^m \sum_{j\neq p} \langle v_s^{(T)}, X[j]\rangle^3$ is $\sigma^3\sqrt{P-1}\sqrt{\sum_{s=1}^m \|v_s^{(T)}\|_2^6}$-subGaussian. Therefore, by applying Lemma J.3, equation (116) is further bounded by:

$$\begin{aligned}\mathscr{L}(f_{W^T}) \leq 2(1-\hat{\mu})\exp\left(-\tilde{\Omega}\left(\frac{\alpha^6}{\beta^6}\right)\frac{1}{\sigma^6\sum_{s=1}^m \|v_s^{(T)}\|_2^6}\right) \\ + 2\hat{\mu}\exp\left(-\frac{\tilde{\Omega}(1)}{\sigma^6\sum_{s=1}^m \|v_s^{(T)}\|_2^6}\right).\end{aligned} \tag{117}$$

Using the fact that $\|v_s^{(T)}\|_2 \leq 1$ in equation (117) finally yields:

$$\mathscr{L}(f_{W^T}) \leq 2(1-\hat{\mu})\exp\left(-\tilde{\Omega}\left(\frac{\alpha^6}{\beta^6\sigma^6}\right)\right) + 2\hat{\mu}\exp\left(-\tilde{\Omega}\left(\frac{1}{\sigma^6}\right)\right). \tag{118}$$

Since $\exp(-\tilde{\Omega}(1/\sigma^6)) \leq 1/\mathrm{poly(d)}$, we obtain the desired result.

□

G EXTENSION TO $\lambda > 0$

Now we discuss how to extend the result to $\lambda > 0$. In our result, since $\lambda = \frac{1}{N\mathrm{poly}(d)}$, we know that before $T = \tilde{\Theta}\left(\frac{1}{\eta\lambda}\right)$ iterations, the weight decay would not affect the learning process and we can show everything similarly.

After iteration T, by Lemma (H.10) and Lemma (I.5), we know that for GD:

$$\nu^{(t)} \leq \tilde{O}(\lambda)$$

and for GD + M:

$$\nu^{(t)} \leq \tilde{O}\left(\lambda/(\beta^2)\right)$$

For GD, we just need to maintain that $c^{(t)} = \tilde{O}(1/\alpha)$ and $\Xi_i^{(t)} = \tilde{\Omega}(1)$. To see this, we know that if $c^{(t)} = \tilde{\Omega}(1/\alpha)$, then

$$c^{(t+1)} \leq (1-\eta\lambda)c^{(t)} + \eta\tilde{O}\left(\nu^{(t)}\frac{\beta^3}{\alpha^2}\right) \leq c^{(t)}$$

To show that $\Xi_i^{(t)} = \tilde{\Omega}(1)$, assuming that $\Xi_i^{(t)} = 1/\mathrm{polylog}(d)$, we know that

$$\Xi_i^{(t+1)} \geq (1-\eta\lambda)\Xi_i^{(t)} + \tilde{\Omega}\left(\eta\frac{1}{N}\right) \geq \Xi_i^{(t)} + \tilde{\Omega}\left(\eta\frac{1}{N}\right)$$

Similarly, for GD + M, since $\nu^{(t)} \leq \tilde{O}\left(\lambda/(\beta^2)\right)$, we know that

$$\nabla\hat{L}(W^{(t)}) \leq \tilde{O}\left(\lambda\alpha^3/(\beta^2)\right)$$

This implies that

$$\|W^{(t+1)} - W^{(t)}\|_2 \leq \tilde{O}\left(\eta\lambda\alpha^3/(\beta^2)\right)$$

We need to show that $c^{(t)} = \tilde{\Omega}(1/\beta)$ and all $|\Xi_{i,j,r}^{(t)}| \leq \tilde{O}(\sigma_0)$. To see this, we know that when $c^{(t)} = \Theta\left(\frac{1}{\beta}\right)$, we know that $c^{(t-t_0)} = \Theta\left(\frac{1}{\beta}\right)$ for every $t_0 \leq \frac{1}{\gamma}$. This implies that

$$c^{(t+1)} \geq c^{(t)} - O\left(\eta\lambda\frac{1}{\beta}\right) + \Omega\left(\frac{\eta}{N}\beta\right) \geq c^{(t)} + \Omega\left(\frac{\eta}{N}\beta\right)$$

On the other hand, for $\Xi_{i,j,r}^{(t)}$ we know that:

$$|\Xi_{i,j,r}^{(t+1)}| \leq (1-\eta\lambda)|\Xi_{i,j,r}^{(t)}| + \tilde{O}\left(\eta\nu^{(t)}\sigma_0^2\right) \leq \tilde{O}(\sigma_0)$$

H TECHNICAL LEMMAS FOR GD

This section presents the technical lemmas needed in Appendix E. These lemmas mainly consists in rewriting of the GD update on the signal and noise components and consequences of such rewriting.

H.1 REWRITING DERIVATIVES FOR GD

Using Induction hypothesis C.1 and Induction hypothesis C.2, we rewrite the sigmoid terms $\ell_i^{(t)}$ when using GD.

Lemma H.1 (Bound on $\mathcal{Z}_1$ derivative for GD)**.** *Let $i \in \mathcal{Z}_1$. We have $\ell_i^{(t)} = \tilde{\Theta}(1)\widehat{\ell}^{(t)}(\alpha)$.*

Proof of Lemma H.1. Let $i \in \mathcal{Z}_1$. Using Induction hypothesis C.1, we bound $\ell_i^{(t)}$ as

$$\frac{1}{1+\exp\left(\alpha^3\sum_{s=1}^m (c_s^{(t)})^3 + \tilde{O}((\sigma\sigma_0\sqrt{d})^3)\right)} \leq \ell_i^{(t)} \leq \frac{1}{1+\exp\left(\alpha^3\sum_{s=1}^m (c_s^{(t)})^3 - \tilde{O}((\sigma\sigma_0\sqrt{d})^3)\right)}$$
$$\iff e^{-\tilde{O}((\sigma\sigma_0\sqrt{d})^3)}\widehat{\ell}^{(t)}(\alpha) \leq \ell_i^{(t)} \leq e^{\tilde{O}((\sigma\sigma_0\sqrt{d})^3)}\widehat{\ell}^{(t)}(\alpha). \tag{119}$$

equation (119) yields the aimed result. □

Lemma H.2 (Bound on $\mathcal{Z}_2$ derivative for GD)**.** *Let $i \in \mathcal{Z}_2$. We have $\ell_i^{(t)} = \tilde{\Theta}(1)\widehat{\ell}^{(t)}(\Xi_i^{(t)})$.*

Proof. Let $i \in \mathcal{Z}_2$. Using Induction hypothesis C.2, we bound $\ell_i^{(t)}$ as

$$\frac{1}{1+\exp\left(\beta^3\tilde{O}(1) + \Xi_i^{(t)}\right)} \leq \ell_i^{(t)} \leq \frac{1}{1+\exp\left(-\beta^3\tilde{O}(\sigma_0^3) + \Xi_i^{(t)}\right)} \tag{120}$$
$$\iff e^{-\tilde{O}(\beta^3)}\widehat{\ell}^{(t)}(\Xi_i^{(t)}) \leq \ell_i^{(t)} \leq e^{\beta^3\tilde{O}(\sigma_0^3)}\widehat{\ell}^{(t)}(\Xi_i^{(t)}). \tag{121}$$

equation (121) yields the aimed result. □

H.2 PROJECTION ONTO THE SIGNAL COMPONENT

Lemma H.3. *We have:*

$$c^{(t+1)} - c^{(t)} \leq \tilde{O}(\eta)\left((1-\hat{\mu})\widehat{\mathcal{L}}^{(t)}(\alpha) + \frac{\tilde{O}(\beta^3)}{N}\sum_{i\in\mathcal{Z}_2}\ell_i^{(t)}\right).$$

Proof. On one hand, we know that

$$c^{(t+1)} - c^{(t)} \leq \sum_{r=1}^{m} |c_r^{(t+1)} - c_r^{(t)}|. \tag{122}$$

On the other hand, from Lemma E.1, we know that:

$$\begin{aligned}\sum_{r=1}^{m} |c_r^{(t+1)} - c_r^{(t)}| \leq \tilde{\Theta}(\eta)(1-\hat{\mu})\alpha \frac{\sum_{r=1}^{m} \alpha^2 (c_r^{(t)})^2}{1 + \exp\left(\sum_{s=1}^{m} \alpha^3 (c_s^{(t)})^3\right)} \\ + \frac{\tilde{\Theta}(\eta\beta^3)}{N} \sum_{i \in \mathcal{Z}_2} \ell_i^{(t)} \sum_{r=1}^{m} (c_r^{(t)})^2.\end{aligned} \tag{123}$$

Using Lemma J.20, we bound the first term in equation (123).

$$\sum_{r=1}^{m} |c_r^{(t+1)} - c_r^{(t)}| \leq \tilde{O}(\eta)(1-\hat{\mu})\alpha \widehat{\mathcal{L}}^{(t)}(\alpha) + \frac{\tilde{\Theta}(\eta\beta^3)}{N} \sum_{i \in \mathcal{Z}_2} \ell_i^{(t)} \sum_{r=1}^{m} (c_r^{(t)})^2. \tag{124}$$

Using Induction hypothesis C.2, we bound $(c_r^{(t)})^2$ in the second term in equation (124).

$$\sum_{r=1}^{m} |c_r^{(t+1)} - c_r^{(t)}| \leq \tilde{O}(\eta)(1-\hat{\mu})\alpha \widehat{\mathcal{L}}^{(t)}(\alpha) + \frac{\tilde{O}(\eta\beta^3)}{N} \sum_{i \in \mathcal{Z}_2} \ell_i^{(t)}. \tag{125}$$

where we used the fact that $m \leq \tilde{O}(1)$. The desired result is obtained by combining equation (122) and equation (125).

□

H.3 Bound on $\mathcal{Z}_1$ derivative

Lemma H.4. *Let $t, \mathscr{T} \in [T]$ such that $\mathscr{T} < t$. Then, the $\mathcal{Z}_1$ derivative is bounded as:*

$$\sum_{\tau=\mathscr{T}}^{t} \nu_1^{(\tau)} \min\{\kappa, \alpha^2 (c^{(\tau)})^2\} \leq \tilde{O}\left(\frac{1}{\eta\alpha^2}\right) + \tilde{O}\left(\frac{\beta^3}{\alpha^2}\right) \sum_{\tau=\mathscr{T}}^{t} \nu_2^{(\tau)}.$$

Proof of Lemma H.4. From Lemma E.4, we know that:

$$c^{(t+1)} \geq c^{(t)} + \tilde{\Theta}(\eta\alpha) \nu_1^{(t)} \min\{\kappa, \alpha^2 (c^{(t)})^2\} \tag{126}$$

Let $\mathscr{T}, t \in [T]$ such that $\mathscr{T} < t$. We now sum up equation (126) for $\tau = \mathscr{T}, \ldots, t$ and get:

$$\sum_{\tau=\mathscr{T}}^{t} \nu_1^{(\tau)} \min\{\kappa, \alpha^2 (c^{(\tau)})^2\} \leq \tilde{O}\left(\frac{1}{\eta\alpha}\right) (c^{(t+1)} - c^{(\mathscr{T})}). \tag{127}$$

We now consider two cases.

Case 1: $t < T_0$. By definition, we know that $c^{(t)} \leq \tilde{O}(1/\alpha)$. Therefore, equation (127) yields:

$$\sum_{\tau=\mathscr{T}}^{t} \nu_1^{(\tau)} \min\{\kappa, \alpha^2 (c^{(\tau)})^2\} \leq \tilde{O}\left(\frac{1}{\eta\alpha^2}\right) \leq \tilde{O}\left(\frac{1}{\eta\alpha}\right). \tag{128}$$

Case 2: $t \in [T_0, T]$. We distinguish two subcases.

- **Subcase 1**: $\mathscr{T} < T_0$. From Lemma 4.3, we know that:

$$c^{(t+1)} \leq \tilde{O}(1/\alpha) + \tilde{O}(\eta\beta^3/\alpha^2) \sum_{\tau=T_0}^{t} \nu_2^{(\tau)}. \tag{129}$$

We can further bound equation (129) as:

$$c^{(t+1)} \leq \tilde{O}(1/\alpha) + \tilde{O}(\eta\beta^3/\alpha^2) \sum_{\tau=\mathscr{T}}^{t} \nu_2^{(\tau)}, \tag{130}$$

which combined with equation (127) implies:

$$\sum_{\tau=\mathscr{T}}^{t} \nu_1^{(\tau)} \min\{\kappa, \alpha^2 (c^{(\tau)})^2\} \leq \tilde{O}\left(\frac{1}{\eta\alpha^2}\right) + \tilde{O}\left(\frac{\beta^3}{\alpha^2}\right) \sum_{\tau=\mathscr{T}}^{t} \nu_2^{(\tau)} \tag{131}$$

– **Subcase 2**: $\mathscr{T} > T_0$. From Lemma 4.3, we know that:

$$c^{(t+1)} \leq \tilde{O}(1/\alpha) + \tilde{O}(\eta\beta^3.\alpha^2) \sum_{\tau=\mathcal{T}}^{t} \nu_2^{(\tau)}, \tag{132}$$

which combined with equation (127) yields equation (131).

We therefore managed to prove that in all the cases, equation (131) holds.

□

H.4 Projection on $\mathcal{Z}_2$ derivative

Lemma H.5. *Let $i \in [N]$, $j \in [P]\backslash\{P(X_i)\}$ and $r \in [m]$. Let $\mathscr{T}, t \in [T]$ such that $\mathscr{T} < t$. Then, the noise update equation (GD-N) satisfies*

$$\left| y_i(\Xi_{i,j,r}^{(t)} - \Xi_{i,j,r}^{(\mathscr{T})}) - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N} \sum_{\tau=\mathscr{T}}^{t-1} \ell_i^{(\tau)} (\Xi_{i,j,r}^{(\tau)})^2 \right| \leq \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right) + \tilde{O}\left(\frac{\eta\beta^3}{\alpha}\right) \sum_{j=\mathscr{T}}^{t-1} \nu_2^{(j)}.$$

Proof of Lemma H.5. Let $i \in [N]$, $j \in [P]\backslash\{P(X_i)\}$ and $r \in [m]$. We set up the following induction hypothesis:

$$\begin{aligned} \left| y_i\Xi_{i,j,r}^{(t)} - y_i\Xi_{i,j,r}^{(\mathscr{T})} - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N} \sum_{\tau=\mathscr{T}}^{t-1} \ell_i^{(\tau)} (\Xi_{i,j,r}^{(\tau)})^2 \right| &\leq \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\left(1 + \frac{\alpha}{\sigma^2 d} + \frac{\alpha\eta}{N}\right) \sum_{\tau=0}^{t-1-\mathscr{T}} \frac{P^\tau}{d^{\tau/2}}\right) \\ &\quad + \tilde{O}\left(\frac{\eta\beta^3}{\alpha^2}\right) \sum_{\tau=0}^{t-1-\mathscr{T}} \frac{P^\tau}{d^{\tau/2}} \sum_{j=\mathscr{T}}^{t-\tau} \nu_2^{(j)}, \end{aligned} \tag{133}$$

Let's first show this hypothesis for $t = \mathscr{T}$. From Lemma E.5, we have:

$$\begin{aligned} \left| y_i(\Xi_{i,j,r}^{(\mathscr{T}+1)} - \Xi_{i,j,r}^{(\mathscr{T})}) - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N} \ell_i^{(\mathscr{T})} (\Xi_{i,j,r}^{(\mathscr{T})})^2 \right| &\leq \frac{\tilde{\Theta}(\eta\sigma^2\sqrt{d})}{N} \sum_{a\in\mathcal{Z}_2} \sum_{k\neq P(X_a)} \ell_a^{(\mathscr{T})} (\Xi_{a,k,r}^{(\mathscr{T})})^2 \\ &\quad + \frac{\tilde{\Theta}(\eta\sigma^2\sqrt{d})}{N} \sum_{a\in\mathcal{Z}_1} \sum_{k\neq P(X_a)} \ell_a^{(\mathscr{T})} (\Xi_{a,k,r}^{(\mathscr{T})})^2. \end{aligned} \tag{134}$$

Now, we apply Induction hypothesis C.3 to bound $(\Xi_{a,k,r}^{(\mathscr{T})})^2$ in equation (134) and obtain:

$$\begin{aligned} \left| y_i\Xi_{i,j,r}^{(\mathscr{T}+1)} - y_i\Xi_{i,j,r}^{(\mathscr{T})} - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N} \ell_i^{(\mathscr{T})} (\Xi_{i,j,r}^{(\mathscr{T})})^2 \right| &\leq \tilde{\Theta}(\eta P\sigma^2\sqrt{d}) \nu_2^{(\mathscr{T})} \min\{\kappa, (c^{(\mathscr{T})})^2\alpha^2\}\alpha \\ &\quad + \tilde{\Theta}(\eta P\sigma^2\sqrt{d}) \nu_1^{(\mathscr{T})} \min\{\kappa, (c^{(\mathscr{T})})^2\alpha^2\}\alpha. \end{aligned} \tag{135}$$

We successively apply Lemma H.4, use $\nu_2^{(\mathscr{T})} \min\{\kappa, (c^{(\mathscr{T})})^2\alpha^2\}\alpha \leq \hat{\mu}\tilde{O}(1) \leq \tilde{O}(\hat{\mu})$ and $\hat{\mu} = \tilde{\Theta}(1/N)$ in equation (135) to finally obtain:

$$\left| y_i \Xi_{i,j,r}^{(\mathscr{T}+1)} - y_i \Xi_{i,j,r}^{(\mathscr{T})} - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N} \ell_i^{(\mathscr{T})} (\Xi_{i,j,r}^{(\mathscr{T})})^2 \right| \leq \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha} \left(1 + \frac{\eta\alpha}{N}\right) \right).$$

Therefore, the induction hypothesis is verified for $t = \mathscr{T}$. Now, assume equation (133) for t. Let's prove the result for $t+1$. We start by summing up the noise update from Lemma E.5 for $\tau = \mathscr{T}, \ldots, t$ which yields:

$$\begin{aligned} \left| y_i (\Xi_{i,j,r}^{(t+1)} - \Xi_{i,j,r}^{(\mathscr{T})}) - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N} \sum_{\tau=\mathscr{T}}^{t} \ell_i^{(\tau)} (\Xi_{i,j,r}^{(\tau)})^2 \right| &\leq \frac{\tilde{\Theta}(\eta\sigma^2\sqrt{d})}{N} \sum_{\tau=\mathscr{T}}^{t-1} \sum_{a\in\mathcal{Z}_2} \ell_a^{(\tau)} \sum_{k\neq P(X_a)} (\Xi_{a,k,r}^{(\tau)})^2 \\ &+ \frac{\tilde{\Theta}(\eta\sigma^2\sqrt{d})}{N} \sum_{a\in\mathcal{Z}_2} \ell_a^{(t)} \sum_{k\neq P(X_a)} (\Xi_{a,k,r}^{(t)})^2 \\ &+ \frac{\tilde{\Theta}(\eta\sigma^2\sqrt{d})}{N} \sum_{\tau=\mathscr{T}}^{t} \sum_{a\in\mathcal{Z}_1} \ell_a^{(\tau)} \sum_{k\neq P(X_a)} (\Xi_{a,k,r}^{(\tau)})^2 \end{aligned} \tag{136}$$

We apply Induction hypothesis C.3 to bound $(\Xi_{a,k,r}^{(t)})^2$ in equation (136) and obtain:

$$\begin{aligned} \left| y_i (\Xi_{i,j,r}^{(t+1)} - \Xi_{i,j,r}^{(\mathscr{T})}) - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N} \sum_{\tau=\mathscr{T}}^{t} \ell_i^{(\tau)} (\Xi_{i,j,r}^{(\tau)})^2 \right| &\leq \frac{\tilde{\Theta}(\eta\sigma^2\sqrt{d})}{N} \sum_{\tau=\mathscr{T}}^{t-1} \sum_{a\in\mathcal{Z}_2} \ell_a^{(\tau)} \sum_{k\neq P(X_a)} (\Xi_{a,k,r}^{(\tau)})^2 \\ &+ \tilde{\Theta}(\eta P\sigma^2\sqrt{d}) \nu_2^{(t)} \alpha \min\{\kappa, (c^{(t)})^2\alpha^2\} \\ &+ \tilde{\Theta}(\eta P\sigma^2\sqrt{d}) \sum_{\tau=\mathscr{T}}^{t} \nu_1^{(\tau)} \alpha \min\{\kappa, (c^{(\tau)})^2\alpha^2\} \end{aligned} \tag{137}$$

Similarly to above, we apply Lemma H.4 to bound $\sum_{\tau=0}^{t} \nu_1^{(\tau)} \alpha \min\{\kappa, (c^{(\tau)})^2\alpha^2\}$. We also use $\nu_2^{(t)} \alpha \min\{\kappa, (c^{(t)})^2\alpha^2\} \leq \tilde{O}(\hat{\mu})$ and $\hat{\mu} = \tilde{\Theta}(1/N)$ in equation (137) and obtain:

$$\begin{aligned} \left| y_i (\Xi_{i,j,r}^{(t+1)} - \Xi_{i,j,r}^{(\mathscr{T})}) - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N} \sum_{\tau=\mathscr{T}}^{t} \ell_i^{(\tau)} (\Xi_{i,j,r}^{(\tau)})^2 \right| &\leq \frac{\tilde{\Theta}(\eta\sigma^2\sqrt{d})}{N} \sum_{\tau=\mathscr{T}}^{t-1} \sum_{a\in\mathcal{Z}_2} \ell_a^{(\tau)} \sum_{k\neq P(X_a)} (\Xi_{a,k,r}^{(\tau)})^2 \\ &+ \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha} \left(1 + \frac{\eta\alpha}{N}\right) \right) \\ &+ \tilde{O}\left(\frac{\eta\beta^3}{\alpha^2} \right) \sum_{j=\mathscr{T}}^{t} \nu_2^{(j)}. \end{aligned} \tag{138}$$

To bound the first term in the right-hand side of equation (138), we use the induction hypothesis equation (133). Plugging this inequality in equation (138) yields:

$$\begin{aligned}\left|y_i(\Xi_{i,j,r}^{(t+1)} - \Xi_{i,j,r}^{(\mathscr{T})}) - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N}\sum_{\tau=\mathscr{T}}^{t}\ell_i^{(\tau)}(\Xi_{i,j,r}^{(\tau)})^2\right| &\leq \frac{1}{\sqrt{d}}\sum_{a\in\mathcal{Z}_2}\sum_{k\neq P(X_k)} y_a(\Xi_{a,k,r}^{(t)} - \Xi_{a,k,r}^{(\mathscr{T})}) \\ &+ \tilde{O}\left(\frac{P^2\sigma^2}{\alpha\sqrt{d}}\left(1+\frac{\alpha}{\sigma^2 d}+\frac{\alpha\eta}{N}\right)\sum_{\tau=0}^{t-1-\mathscr{T}}\frac{P^\tau}{d^{\tau/2}}\right) \\ &+ \frac{P}{\sqrt{d}}\tilde{O}\left(\frac{\eta\beta^3}{\alpha^2}\right)\sum_{\tau=0}^{t-1-\mathscr{T}}\frac{P^\tau}{d^{\tau/2}}\sum_{j=\mathscr{T}}^{t-1-\tau}\nu_2^{(j)} \\ &+ \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\left(1+\frac{\eta\alpha}{N}\right)\right) \\ &+ \tilde{O}\left(\frac{\eta\beta^3}{\alpha^2}\right)\sum_{j=\mathscr{T}}^{t}\nu_2^{(j)}.\end{aligned} \tag{139}$$

Now, we apply Induction hypothesis C.1 to have $y_a(\Xi_{a,k,r}^{(t)} - \Xi_{a,k,r}^{(0)}) \leq \tilde{O}(1)$ in equation (139) and therefore,

$$\begin{aligned}\left|y_i\Xi_{i,j,r}^{(t+1)} - y_i\Xi_{i,j,r}^{(\mathscr{T})} - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N}\sum_{\tau=\mathscr{T}}^{t}\ell_i^{(\tau)}(\Xi_{i,j,r}^{(\tau)})^2\right| &\leq \frac{\tilde{O}(P)}{\sqrt{d}} \\ &+ \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\left(1+\frac{\alpha}{\sigma^2 d}+\frac{\alpha\eta}{N}\right)\sum_{\tau=1}^{t-\mathscr{T}}\frac{P^\tau}{d^{\tau/2}}\right) \\ &+ \tilde{O}\left(\frac{\eta\beta^3}{\alpha^2}\right)\sum_{\tau=1}^{t-\mathscr{T}}\frac{P^\tau}{d^{\tau/2}}\sum_{j=\mathscr{T}}^{t-\tau}\nu_2^{(j)} \\ &+ \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\left(1+\frac{\eta\alpha}{N}\right)\right) \\ &+ \tilde{O}\left(\frac{\eta\beta^3}{\alpha^2}\right)\sum_{j=\mathscr{T}}^{t}\nu_2^{(j)}.\end{aligned} \tag{140}$$

By rearranging the terms, we finally have:

$$\begin{aligned}\left|y_i\Xi_{i,j,r}^{(t+1)} - y_i\Xi_{i,j,r}^{(\mathscr{T})} - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N}\sum_{\tau=\mathscr{T}}^{t}\ell_i^{(\tau)}(\Xi_{i,j,r}^{(\tau)})^2\right| &\leq \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\left(1+\frac{\alpha}{\sigma^2 d}+\frac{\alpha\eta}{N}\right)\sum_{\tau=0}^{t-\mathscr{T}}\frac{P^\tau}{d^{\tau/2}}\right) \\ &+ \tilde{O}\left(\frac{\eta\beta^3}{\alpha^2}\right)\sum_{\tau=0}^{t-\mathscr{T}}\frac{P^\tau}{d^{\tau/2}}\sum_{j=\mathscr{T}}^{t-\tau}\nu_2^{(j)},\end{aligned} \tag{141}$$

which proves the induction hypothesis for $t+1$.

Now, let's simplify the sum terms in equation (133). Since $P \ll \sqrt{d}$, by definition of a geometric sequence, we have:

$$\sum_{\tau=0}^{t-\mathscr{T}}\frac{P^\tau}{d^{\tau/2}} \leq \frac{1}{1-\frac{P}{\sqrt{d}}} \leq O(1). \tag{142}$$

Plugging equation (142) in equation (133) yields

$$\left|y_i(\Xi_{i,j,r}^{(t)} - \Xi_{i,j,r}^{(\mathscr{T})}) - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N}\sum_{\tau=\mathscr{T}}^{t}\ell_i^{(\tau)}(\Xi_{i,j,r}^{(\tau)})^2\right| \leq \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right) + \tilde{O}\left(\frac{\eta\beta^3}{\alpha^2}\right)\sum_{\tau=0}^{t-1-\mathscr{T}}\frac{P^\tau}{d^{\tau/2}}\sum_{j=\mathscr{T}}^{t-1-\tau}\nu_2^{(j)}. \tag{143}$$

Now, let's simplify the second sum term in equation (143). Indeed, we have:

$$\sum_{\tau=0}^{t-1-\mathscr{T}}\frac{P^\tau}{d^{\tau/2}}\sum_{j=\mathscr{T}}^{t-1-\tau}\nu_2^{(j)} \leq \sum_{\tau=0}^{t-1-\mathscr{T}}\frac{P^\tau}{d^{\tau/2}}\sum_{j=\mathscr{T}}^{t-1}\nu_2^{(j)} \leq O(1)\sum_{j=\mathscr{T}}^{t-1}\nu_2^{(j)}, \tag{144}$$

where we used equation (142) in the last inequality. Plugging equation (144) in equation (143) gives the final result. □

After T_1 iterations, we prove with Lemma 4.4 that there exists j, r $\Xi_{i,j,r}^{(t)}$ that becomes very large. However, we would like to claim that $\ell_i^{(\tau)}(\Xi_i^{(\tau)})$ stays controlled thanks to the sigmoid. We would like to have a noise update that takes this into account.

Lemma H.6 (Noise update at late iterations)**.** *Let $i \in [N]$, $j \in [P]\backslash\{P(X_i)\}$ and $r \in [m]$. Let $\mathscr{T}, t \in [T]$ such that $\mathscr{T} < t$. Then, the noise update equation (GD-N) satisfies*

$$\left|y_i(\Xi_{i,j,r}^{(t)} - \Xi_{i,j,r}^{(\mathscr{T})}) - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N}\sum_{\tau=\mathscr{T}}^{t-1}\ell_i^{(\tau)}\min\{\kappa, (\Xi_{i,j,r}^{(\tau)})^2\}\right| \leq \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right) + \tilde{O}\left(\frac{\eta\beta^3}{\alpha^2}\right)\sum_{j=\mathscr{T}}^{t-1}\nu_2^{(j)}.$$

Proof of Lemma H.6. From Lemma H.5, we know that

$$\left|y_i(\Xi_{i,j,r}^{(t)} - \Xi_{i,j,r}^{(\mathscr{T})}) - \frac{\tilde{\Theta}(\eta\sigma^2 d)}{N}\sum_{\tau=\mathscr{T}}^{t-1}\ell_i^{(\tau)}(\Xi_{i,j,r}^{(\tau)})^2\right| \leq \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right) + \tilde{O}\left(\frac{\eta\beta^3}{\alpha^2}\right)\sum_{j=\mathscr{T}}^{t-1}\nu_2^{(j)}. \tag{145}$$

Using Remark 1, we know that a sufficient condition to have $\widehat{\ell^{(\tau)}}(\Xi_i^{(t)}$ is $(\Xi_{i,j,r}^{(\tau)})^2 \geq \kappa \geq \tilde{\Omega}(1)$. Therefore, we can replace $\widehat{\ell^{(t)}}(\Xi_i^{(t)})(\Xi_{i,j,r}^{(\tau)})^2 = \min\{\kappa, (\Xi_{i,j,r}^{(\tau)})^2\}$. Plugging this equality in equation (145) yields the aimed result. □

Lemma H.7. *Let $T_1 = \tilde{O}\left(\frac{N}{\sigma_0\sigma\sqrt{d}\sigma^2 d}\right)$. For $t \in [T_1, T]$, we have $\frac{1}{N}\sum_{\tau=0}^{t}\sum_{i\in\mathcal{Z}_2}\ell_i^{(\tau)}\min\{\kappa, (\Xi_{i,j,r}^{(\tau)})^2\} \leq \tilde{O}\left(\frac{1}{\eta}\right)$.*

Proof of Lemma H.7. From Lemma E.7, we know that:

$$\sum_{\tau=T_1}^{t}\nu_2^{(\tau)} \leq \tilde{O}\left(\frac{1}{\eta\sigma_0}\right). \tag{146}$$

On the other hand we know from Lemma H.6 that:

$$\frac{\tilde{\Theta}(\eta\sigma^2 d)}{N}\sum_{\tau=0}^{T_1-1}\sum_{i\in\mathcal{Z}_2}\ell_i^{(\tau)}\min\{\kappa, (\Xi_{i,j,r}^{(\tau)})^2\} \leq y_i(\Xi_{i,j,r}^{(T_1)} - \Xi_{i,j,r}^{(0)}) + \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right) + \tilde{O}\left(\frac{\eta\hat{\mu}\beta^3}{\alpha}\right)T_1. \tag{147}$$

Besides, we have: $\tilde{O}\left(\frac{\eta\hat{\mu}\beta^3}{\alpha}\right)T_1 \leq \tilde{O}\left(\frac{\beta^3 N\hat{\mu}}{\alpha\sigma_0\sigma\sqrt{d}\sigma^2 d}\right) \leq \tilde{O}\left(\frac{\beta^3}{\alpha\sigma_0\sigma\sqrt{d}\sigma^2 d}\right) \leq \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right)$, where we used $\hat{\mu} = \tilde{\Theta}(1/N)$ in the penultimate inequality. Plugging this inequality yields

$$\frac{\tilde{\Theta}(\eta\sigma^2 d)}{N}\sum_{\tau=0}^{T_1-1}\sum_{i\in\mathcal{Z}_2}\ell_i^{(\tau)}\min\{\kappa, (\Xi_{i,j,r}^{(\tau)})^2\} \leq y_i(\Xi_{i,j,r}^{(T_1)} - \Xi_{i,j,r}^{(0)}) + \tilde{O}\left(\frac{P\sigma^2\sqrt{d}}{\alpha}\right). \tag{148}$$

By applying Induction hypothesis C.1, equation (148) is eventually bounded as:

$$\frac{1}{N}\sum_{\tau=0}^{t}\sum_{i\in\mathcal{Z}_2}\ell_i^{(\tau)}\min\{\kappa,(\Xi_{i,j,r}^{(\tau)})^2\}\leq\tilde{O}\left(\frac{1}{\eta\sigma^2 d}\right)+\tilde{O}\left(\frac{P}{\eta\alpha\sqrt{d}}\right)\leq\tilde{O}\left(\frac{1}{\eta}\right). \tag{149}$$

By combining equation (44) and equation (149) we deduce that for all $j\in[P]\backslash\{P(X_i)\}$ and $r\in[m]$:

$$\begin{aligned}\frac{1}{N}\sum_{\tau=0}^{t}\sum_{i\in\mathcal{Z}_2}\ell_i^{(\tau)}\min\{\kappa,(\Xi_{i,j,r}^{(\tau)})^2\}&=\frac{1}{N}\sum_{\tau=0}^{T_1}\sum_{i\in\mathcal{Z}_2}\ell_i^{(\tau)}\min\{\kappa,(\Xi_{i,j,r}^{(\tau)})^2\}\\&\quad+\frac{1}{N}\sum_{\tau=T_1}^{t}\sum_{i\in\mathcal{Z}_2}\ell_i^{(\tau)}\min\{\kappa,(\Xi_{i,j,r}^{(\tau)})^2\}\\&\leq\tilde{O}(1)\tilde{O}\left(\frac{1}{\eta}\right)+\tilde{O}\left(\frac{1}{\eta}\right)\\&\leq\tilde{O}\left(\frac{1}{\eta}\right).\end{aligned} \tag{150}$$

□

H.5 Projection on normalized noise

Lemma H.8 (Gradient on the normalized noise)**.** *For $r\in[m]$, the gradient of the loss $\widehat{L}(W^{(t)})$ projected on the normalized noise χ satisfies with probability $1-o(1)$ for $r\in[m]$:*

$$-\mathrm{G}_r^{(t)}\geq\frac{\tilde{\Theta}(\sigma\sqrt{d})}{N}\sum_{i\in\mathcal{Z}_2}\ell_i^{(t)}\sum_{j\neq P(X_i)}(\Xi_{i,j,r}^{(t)})^2-\frac{\tilde{O}(\sigma)}{N}\sum_{i\in\mathcal{Z}_1}\sum_{j\neq P(X_i)}\ell_i^{(t)}(\Xi_{i,j,r}^{(t)})^2.$$

Proof of Lemma H.8. Projecting the gradient (given by Lemma D.1) on χ yields:

$$\begin{aligned}-\mathrm{G}_r^{(t)}&=\frac{3}{N^2}\sum_{i\in\mathcal{Z}_2}\sum_{j\neq P(X_i)}\ell_i^{(t)}(\Xi_{i,j,r}^{(t)})^2\frac{\|X_i[j]\|_2^2}{\|\frac{1}{N}\sum_{b\in\mathcal{Z}_2}\sum_{l\neq P(X_i)}X_b[l]\|_2}\\&\quad+\frac{3}{N^2}\sum_{i\in\mathcal{Z}_2}\ell_i^{(t)}\sum_{j\neq P(X_i)}\sum_{\substack{k\neq P(X_i)\\k\neq j}}(\Xi_{i,k,r}^{(t)})^2\left\langle X_i[k],\frac{X_i[j]}{\|\frac{1}{N}\sum_{b\in\mathcal{Z}_2}\sum_{l\neq P(X_i)}X_b[l]\|_2}\right\rangle\\&\quad+\frac{3}{N^2}\sum_{i\in\mathcal{Z}_2}\sum_{\substack{a\in\mathcal{Z}_2\\a\neq i}}\ell_a^{(t)}\sum_{k\neq P(X_a)}(\Xi_{a,k,r}^{(t)})^2\sum_{j\neq P(X_i)}\left\langle X_a[k],\frac{X_i[j]}{\|\frac{1}{N}\sum_{b\in\mathcal{Z}_2}\sum_{l\neq P(X_i)}X_b[l]\|_2}\right\rangle\\&\quad+\frac{3}{N}\sum_{a\in\mathcal{Z}_1}\sum_{k\neq P(X_a)}\ell_a^{(t)}(\Xi_{a,k,r}^{(t)})^2\left\langle X_a[k],\frac{\frac{1}{N}\sum_{i\in\mathcal{Z}_2}\sum_{j\neq P(X_i)}X_i[j]}{\|\frac{1}{N}\sum_{b\in\mathcal{Z}_2}\sum_{l\neq P(X_i)}X_b[l]\|_2}\right\rangle.\end{aligned} \tag{151}$$

We further bound equation (151) as:

$$\begin{aligned}&\left|\mathrm{G}_r^{(t)}+\frac{3}{N^2}\sum_{i\in\mathcal{Z}_2}\sum_{j\neq P(X_i)}\ell_i^{(t)}(\Xi_{i,j,r}^{(t)})^2\frac{\|X_i[j]\|_2^2}{\|\frac{1}{N}\sum_{b\in\mathcal{Z}_2}\sum_{l\neq P(X_i)}X_b[l]\|_2}\right.\\&\left.-\frac{3}{N^2}\sum_{i\in\mathcal{Z}_2}\sum_{a\in\mathcal{Z}_2}\ell_a^{(t)}\sum_{j\neq P(X_i)}\sum_{k\neq P(X_a)}(\Xi_{a,k,r}^{(t)})^2\left|\left\langle X_a[k],\frac{X_i[j]}{\|\frac{1}{N}\sum_{b\in\mathcal{Z}_2}\sum_{l\neq P(X_i)}X_b[l]\|_2}\right\rangle\right|\right|\\&\leq\frac{3}{N}\sum_{a\in\mathcal{Z}_1}\sum_{k\neq P(X_a)}\ell_a^{(t)}(\Xi_{a,k,r}^{(t)})^2\left|\left\langle X_a[k],\frac{\frac{1}{N}\sum_{i\in\mathcal{Z}_2}\sum_{j\neq P(X_i)}X_i[j]}{\|\frac{1}{N}\sum_{b\in\mathcal{Z}_2}\sum_{l\neq P(X_i)}X_b[l]\|_2}\right\rangle\right|.\end{aligned} \tag{152}$$

Since $\frac{\frac{1}{N}\sum_{i\in\mathcal{Z}_2}\sum_{j\neq P(X_i)}X_i[j]}{\|\frac{1}{N}\sum_{b\in\mathcal{Z}_2}\sum_{l\neq P(X_i)}X_b[l]\|_2}$ is a unit Gaussian vector, using Lemma J.8, we bound the right-hand side of equation (152) with probability $1-o(1)$, as:

$$\begin{aligned}&\left|\mathrm{G}_r^{(t)}+\frac{3}{N^2}\sum_{i\in\mathcal{Z}_2}\sum_{j\neq P(X_i)}\ell_i^{(t)}(\Xi_{i,j,r}^{(t)})^2\frac{\|X_i[j]\|_2^2}{\|\frac{1}{N}\sum_{b\in\mathcal{Z}_2}\sum_{l\neq P(X_i)}X_b[l]\|_2}\right.\\&\left.-\frac{3}{N^2}\sum_{i\in\mathcal{Z}_2}\sum_{a\in\mathcal{Z}_2}\ell_a^{(t)}\sum_{j\neq P(X_i)}\sum_{k\neq P(X_a)}(\Xi_{a,k,r}^{(t)})^2\left|\left\langle X_a[k],\frac{X_i[j]}{\|\frac{1}{N}\sum_{b\in\mathcal{Z}_2}\sum_{l\neq P(X_i)}X_b[l]\|_2}\right\rangle\right|\right|\\&\leq\frac{\sigma}{N}\sum_{a\in\mathcal{Z}_1}\sum_{k\neq P(X_a)}\ell_a^{(t)}(\Xi_{a,k,r}^{(t)})^2.\end{aligned}\tag{153}$$

Now, using Lemma Lemma J.10 , we can further lower bound the left-hand side of equation (153) as:

$$\begin{aligned}&\left|\mathrm{G}_r^{(t)}+\frac{3}{N^2}\sum_{i\in\mathcal{Z}_2}\sum_{j\neq P(X_i)}\ell_i^{(t)}(\Xi_{i,j,r}^{(t)})^2\frac{\|X_i[j]\|_2^2}{\|\frac{1}{N}\sum_{b\in\mathcal{Z}_2}\sum_{l\neq P(X_i)}X_b[l]\|_2}\right.\\&\left.-\frac{\tilde{\Theta}(P)}{\sqrt{d}N^2}\sum_{a\in\mathcal{Z}_2}\ell_a^{(t)}\sum_{k\neq P(X_a)}(\Xi_{a,k,r}^{(t)})^2\frac{\|X_a[k]\|_2^2}{\|\frac{1}{N}\sum_{b\in\mathcal{Z}_2}\sum_{l\neq P(X_i)}X_b[l]\|_2}\right|\\&\leq\frac{\sigma}{N}\sum_{a\in\mathcal{Z}_1}\sum_{k\neq P(X_a)}\ell_a^{(t)}(\Xi_{a,k,r}^{(t)})^2.\end{aligned}\tag{154}$$

Rewriting equation (154) yields:

$$\begin{aligned}&\left|\mathrm{G}_r^{(t)}+\frac{\Theta(1)}{N^2}\sum_{i\in\mathcal{Z}_2}\sum_{j\neq P(X_i)}\ell_i^{(t)}(\Xi_{i,j,r}^{(t)})^2\frac{\|X_i[j]\|_2^2}{\|\frac{1}{N}\sum_{b\in\mathcal{Z}_2}\sum_{l\neq P(X_i)}X_b[l]\|_2}\right|\\&\leq\frac{\sigma}{N}\sum_{a\in\mathcal{Z}_1}\sum_{k\neq P(X_a)}\ell_a^{(t)}(\Xi_{a,k,r}^{(t)})^2.\end{aligned}\tag{155}$$

Remark that $\frac{1}{N}\sum_{b\in\mathcal{Z}_2}\sum_{l\neq P(X_i)}X_b[l]\sim\mathcal{N}(0,\frac{\hat{\mu}P}{N}\sigma^2)$. By applying Lemma J.9, we have:

$$\frac{1}{N}\frac{\|X_i[j]\|_2^2}{\|\frac{1}{N}\sum_{b\in\mathcal{Z}_2}\sum_{l\neq P(X_i)}X_b[l]\|_2}=\frac{1}{N}\tilde{\Theta}\left(\sigma\sqrt{\frac{dN}{\hat{\mu}P}}\right)=\tilde{\Theta}\left(\sigma\sqrt{\frac{d}{\hat{\mu}NP}}\right)=\tilde{\Theta}(\sigma\sqrt{d}),\tag{156}$$

where we used $P=\tilde{\Theta}(1)$ and $\hat{\mu}N=\tilde{\Theta}(1)$ in the last equality of equation (156). Plugging this in equation (155) yields the desired result. □

H.6 Convergence rate of the training loss using GD

In this section, we prove that when using GD, the training loss converges sublinearly in our setting.

H.6.1 Convergence after learning $\mathcal{Z}_1$ ($t\in[T_0,T]$)

Lemma H.9 (Convergence rate of the $\mathcal{Z}_1$ loss)**.** *Let $t\in[T_0,T]$. Run GD with learning rate $\eta\in(0,1/L)$ for t iterations. Then, the $\mathcal{Z}_1$ loss sublinearly converges to zero as:*

$$(1-\hat{\mu})\widehat{\mathcal{L}}^{(t)}(\alpha)\leq\frac{\tilde{O}(1)}{\eta\alpha^2(t-T_0+1)}.$$

Proof of Lemma H.9. Let $t\in[T_0,T]$. From Lemma E.1, we know that the signal update is lower bounded as:

$$c^{(t+1)}\geq c^{(t)}+\Theta(\eta\alpha)(1-\hat{\mu})\widehat{\ell}^{(t)}(\alpha)(\alpha c^{(t)})^2.\tag{157}$$

From Lemma 4.1, we know that $c^{(t)} \geq \tilde{\Omega}(1/\alpha)$. Thus, we simplify equation (157) as:

$$c^{(t+1)} \geq c^{(t)} + \tilde{\Omega}(\eta\alpha)(1-\hat{\mu})\widehat{\ell}^{(t)}(\alpha). \tag{158}$$

Since $\alpha^3 \sum_{r=1}^m (c_r^{(t)})^3 \geq \tilde{\Omega}(1/\alpha) - m\tilde{O}(\sigma_0) \geq \tilde{\Omega}(1/\alpha) > 0$, we can apply Lemma J.22 and obtain:

$$c^{(t+1)} \geq c^{(t)} + \tilde{\Omega}(\eta\alpha)(1-\hat{\mu})\widehat{\mathcal{L}}^{(t)}(\alpha). \tag{159}$$

Let's now assume by contradiction that for $t \in [T_0, T]$, we have:

$$(1-\hat{\mu})\widehat{\mathcal{L}}^{(t)}(\alpha) > \frac{\tilde{O}(1)}{\eta\alpha^2(t-(T_0-1))}. \tag{160}$$

From the equation (GDM-S) update, we know that $c_r^{(\tau)}$ is a non-decreasing sequence which implies that $\sum_{r=1}^m (\alpha c_r^{(\tau)})^3$ is also non-decreasing. Since $x \mapsto \log(1+\exp(-x))$ is non-increasing, this implies that for $s \leq t$, we have:

$$\frac{\tilde{O}(1)}{\eta\alpha^2(t-T_0)} < (1-\hat{\mu})\widehat{\mathcal{L}}^{(t)}(\alpha) \leq (1-\hat{\mu})\widehat{\mathcal{L}}^{(s)}(\alpha). \tag{161}$$

Plugging equation (161) in the update equation (159) yields for $s \in [T_0, t]$:

$$c^{(s+1)} > c^{(s)} + \frac{\tilde{O}(1)}{\alpha(t-(T_0-1))}. \tag{162}$$

Let $t \in [T_0, T]$. We now sum equation (162) for $s = T_0, \ldots, t-1$ and obtain:

$$c^{(t)} > c^{(T_0)} + \frac{\tilde{O}(1)(t-T_0)}{\alpha(t-(T_0-1))} > \frac{\tilde{O}(1)}{\alpha}, \tag{163}$$

where we used the fact that $c^{(T_0)} \geq \tilde{\Omega}(1/\alpha) \geq 0$ (Lemma 4.1) in the last inequality. Let's now show that equation (163) implies that equation (161) is a contradiction. Indeed, we have:

$$\begin{aligned}
&\eta\alpha^2(t-(T_0-1))(1-\hat{\mu})\widehat{\mathcal{L}}^{(t)}(\alpha) \\
\leq &\eta\alpha^2 T(1-\hat{\mu})\log\left(1+\exp(-(\alpha c^{(t)})^3 - \sum_{r\neq r_{\max}} (\alpha c_r^{(t)})^3\right) \\
\leq &\eta\alpha^2 T(1-\hat{\mu})\log\left(1+\exp(-\tilde{O}(1)\right),
\end{aligned} \tag{164}$$

where we used $\sum_{r\neq r_{\max}} (c_r^{(t)})^3 \geq -m\tilde{O}(\sigma_0^3)$ along with equation (163) in equation (164). We now apply Lemma J.22 in equation (164) and obtain:

$$\eta\alpha^2(t-(T_0-1))(1-\hat{\mu})\widehat{\mathcal{L}}^{(t)}(\alpha) \leq \frac{(1-\hat{\mu})\eta\alpha^2 T}{1+\exp(\tilde{O}(1))}. \tag{165}$$

Since $T \leq d^{O(\log d)}/\eta$, $N = \Theta(1)\mathrm{poly}(d)$, $\alpha = d^{0.49}$, we finally have:

$$\eta\alpha^2(t-(T_0-1))(1-\hat{\mu})\widehat{\mathcal{L}}^{(t)}(\alpha) \leq \frac{\tilde{\Theta}(d^{0.98})\exp(O(\log^2 d))}{1+\exp(\mathrm{polylog}(d))} = o(1) < \tilde{O}(1), \tag{166}$$

which contradicts equation (160). □

H.6.2 Convergence at late stages ($t \in [T_1, T]$)

Lemma H.10 (Convergence rate of the loss)**.** *Let $t \in [T_1, T]$. Run GD with learning rate $\eta \in (0, 1/L)$ for t iterations. Then, the loss sublinearly converges to zero as:*

$$\widehat{L}(W^{(t)}) \leq \frac{\tilde{\Theta}(1)}{\eta(t-T_1+1)}.$$

Proof of Lemma H.10. We first apply the classical descent lemma for smooth functions (Lemma J.18). Since $\widehat{L}(W)$ is smooth, we have:

$$\widehat{L}(W^{(t+1)}) \leq \widehat{L}(W^{(t)}) - \frac{\eta}{2}\|\nabla \widehat{L}(W^{(t)})\|_2^2 = \widehat{L}(W^{(t)}) - \frac{\eta}{2}\sum_{r=1}^{m}\|\nabla_{w_r}\widehat{L}(W^{(t)})\|_2^2. \tag{167}$$

Lemma H.11 provides a lower bound on the gradient. We plug it in equation (167) and get:

$$\widehat{L}(W^{(t+1)}) \leq \widehat{L}(W^{(t)}) - \tilde{\Omega}(\eta)\widehat{L}(W^{(t)})^2. \tag{168}$$

Applying Lemma J.19 to equation (168) yields the aimed result. □

To obtain the convergence rate in Lemma H.10, we used the following auxiliary lemma.

Lemma H.11 (Bound on the gradient for GD)**.** *Let $t \in [T_1, T]$. Run GD for t iterations. Then, the norm of gradient is lower bounded as follows:*

$$\sum_{r=1}^{m}\|\nabla_{w_r}\widehat{L}(W^{(t)})\|_2^2 \geq \tilde{\Omega}(1)\widehat{L}(W^{(t)})^2.$$

Proof of Lemma H.11. Let $t \in [T_1, T]$. To obtain the lower bound, we project the gradient on the the signal and on the noise.

Projection on the signal. Since $\|w^*\|_2 = 1$, we lower bound $\|\nabla_{w_r}\widehat{L}(W^{(t)})\|_2^2$ as

$$\|\nabla_{w_r}\widehat{L}(W^{(t)})\|_2^2 \geq \langle \nabla_{w_r}\widehat{L}(W^{(t)}), w^*\rangle^2 = (\mathscr{G}_r^{(t)})^2. \tag{169}$$

By successively applying Lemma D.2 and Lemma H.1, $(\mathscr{G}_r^{(t)})^2$ is lower bounded as

$$(\mathscr{G}_r^{(t)})^2 \geq \left(\frac{\alpha^3}{N}\sum_{i\in\mathcal{Z}_1}\ell_i^{(t)}(c_r^{(t)})^2\right)^2 \geq \Omega(1)\left(\alpha^3(1-\hat{\mu})\widehat{\ell}^{(t)}(\alpha)(c_r^{(t)})^2\right)^2. \tag{170}$$

Combining equation (169) and equation (170) yields:

$$\|\nabla_{w_r}\widehat{L}(W^{(t)})\|_2^2 \geq \Omega(1)\left(\alpha^3(1-\hat{\mu})\widehat{\ell}^{(t)}(\alpha)(c_r^{(t)})^2\right)^2. \tag{171}$$

Projection on the noise. For a fixed $i \in \mathcal{Z}_2$ and $j \in [P]\backslash\{P(X_i)\}$, we know that $\|\nabla_{w_r}\widehat{L}(W^{(t)})\|_2^2$ is lower bounded as

$$\|\nabla_{w_r}\widehat{L}(W^{(t)})\|_2^2 \geq \left\langle \nabla_{w_r}\widehat{L}(W^{(t)}), \frac{\frac{1}{N}\sum_{i\in\mathcal{Z}_2}\sum_{j\neq P(X_i)}X_i[j]}{\|\frac{1}{N}\sum_{i\in\mathcal{Z}_2}\sum_{j\neq P(X_i)}X_i[j]\|_2}\right\rangle^2 = (\mathrm{G}_r^{(t)})^2. \tag{172}$$

On the other hand, by Lemma H.8, we lower bound $\mathrm{G}_r^{(t)}$ term with probability $1 - o(1)$ as:

$$(\mathrm{G}_r^{(t)})^2 \geq \left(\frac{\tilde{\Omega}(\sigma\sqrt{d})}{N}\sum_{i\in\mathcal{Z}_2}\sum_{j\neq P(X_i)}\ell_i^{(t)}(\Xi_{i,j,r}^{(t)})^2 - \frac{\tilde{O}(\sigma)}{N}\sum_{i\in\mathcal{Z}_1}\sum_{j\neq P(X_i)}\ell_i^{(t)}(\Xi_{i,j,r}^{(t)})^2\right)^2 \tag{173}$$

Gathering the bounds. Combining equation (169), equation (172), equation (170) and equation (173) and using $2a^2 + 2b^2 \geq (a+b)^2$, we thus bound $\|\nabla_{w_r}\widehat{L}(W^{(t)})\|_2^2$ as:

$$\begin{aligned}\|\nabla_{w_r}\widehat{L}(W^{(t)})\|_2^2 \geq \Bigg(&\frac{\alpha + \tilde{O}(\sigma)}{N}\sum_{i\in\mathcal{Z}_1}\ell_i^{(t)}\alpha^2(c_r^{(t)})^2 \\ &+\frac{\tilde{\Omega}(\sigma\sqrt{d})}{N}\sum_{i\in\mathcal{Z}_2}\sum_{j\neq P(X_i)}\ell_i^{(t)}(\Xi_{i,j,r}^{(t)})^2 \\ &-\frac{\tilde{O}(\sigma)}{N}\sum_{i\in\mathcal{Z}_1}\sum_{j\neq P(X_i)}\ell_i^{(t)}\left((\alpha^2(c_r^{(t)})^2 + (\Xi_{i,j,r}^{(t)})^2\right)\Bigg)^2.\end{aligned} \tag{174}$$

We now sum up equation (174) for $r = 1, \dots, m$ and apply Cauchy-Schwarz inequality to get:

$$\sum_{r=1}^{m} \|\nabla_{w_r} \widehat{L}(W^{(t)})\|_2^2 \geq \frac{1}{m} \Bigg(\frac{\alpha + \tilde{O}(\sigma)}{N} \sum_{r=1}^{m} \ell_i^{(t)}(\alpha) \alpha^2 (c_r^{(t)})^2 + \frac{\tilde{\Omega}(\sigma\sqrt{d})}{N} \sum_{i \in \mathcal{Z}_2} \sum_{r=1}^{m} \sum_{j \neq P(X_i)} \ell_i^{(t)} (\Xi_{i,j,r}^{(t)})^2 - \frac{\tilde{O}(\sigma)}{N} \sum_{i \in \mathcal{Z}_1} \sum_{r=1}^{m} \sum_{j \neq P(X_i)} \ell_i^{(t)} \left((\alpha^2 (c_r^{(t)})^2 + (\Xi_{i,j,r}^{(t)})^2 \right) \Bigg)^2 . \tag{175}$$

We apply Lemma H.1 to further lower bound equation (175) and get:

$$\sum_{r=1}^{m} \|\nabla_{w_r} \widehat{L}(W^{(t)})\|_2^2 \geq \Omega\left(\frac{1}{m}\right) \Bigg((\alpha + \tilde{O}(\sigma))(1 - \hat{\mu}) \sum_{r=1}^{m} \widehat{\ell}^{(t)}(\alpha) \alpha^2 (c_r^{(t)})^2 + \frac{\tilde{\Omega}(\sigma\sqrt{d})}{N} \sum_{i \in \mathcal{Z}_2} \sum_{r=1}^{m} \sum_{j \neq P(X_i)} \ell_i^{(t)} (\Xi_{i,j,r}^{(t)})^2 - \frac{\tilde{O}(\sigma)}{N} \sum_{i \in \mathcal{Z}_1} \sum_{r=1}^{m} \sum_{j \neq P(X_i)} \ell_i^{(t)} \left((\alpha^2 (c_r^{(t)})^2 + (\Xi_{i,j,r}^{(t)})^2 \right) \Bigg)^2 . \tag{176}$$

Bound the gradient terms by the loss. Using Lemma H.12, Lemma H.13 and Lemma H.14 we have:

$$(\alpha + \tilde{O}(\sigma))(1 - \hat{\mu}) \sum_{r=1}^{m} \widehat{\ell}^{(t)}(\alpha) \alpha^2 (c_r^{(t)})^2 \geq \tilde{\Omega}(\alpha + \tilde{O}(\sigma)) \mathcal{L}^{(t)}(\alpha), \tag{177}$$

$$\frac{\tilde{O}(\sigma)}{N} \sum_{i \in \mathcal{Z}_1} \sum_{r=1}^{m} \sum_{j \neq P(X_i)} \ell_i^{(t)} \left((\alpha^2 (c_r^{(t)})^2 + (\Xi_{i,j,r}^{(t)})^2 \right) \leq \tilde{O}(\sigma)(1 - \hat{\mu}) \mathcal{L}^{(t)}(\alpha), \tag{178}$$

$$\frac{\tilde{\Omega}(\sigma\sqrt{d})}{N} \sum_{i \in \mathcal{Z}_2} \sum_{r=1}^{m} \sum_{j \neq P(X_i)} \ell_i^{(t)} (\Xi_{i,j,r}^{(t)})^2 \geq \frac{\tilde{\Omega}(\sigma\sqrt{d})}{N} \sum_{i \in \mathcal{Z}_2} \mathcal{L}^{(t)}(\Xi_i^{(t)}). \tag{179}$$

Plugging equation (177), equation (178) and equation (179) in equation (176) yields:

$$\begin{aligned} \sum_{r=1}^{m} \|\nabla_{w_r} \widehat{L}(W^{(t)})\|_2^2 &\geq \Omega\left(\frac{1}{m}\right) \Bigg((\alpha + \tilde{O}(\sigma))(1 - \hat{\mu}) \mathcal{L}^{(t)}(\alpha) + \frac{\tilde{\Omega}(\sigma\sqrt{d})}{N} \sum_{i \in \mathcal{Z}_2} \mathcal{L}^{(t)}(\Xi_i^{(t)}) - (1 - \hat{\mu}) \tilde{O}(\sigma) \mathcal{L}^{(t)}(\alpha) \Bigg)^2 \\ &\geq \tilde{\Omega}(1) \left((1 - \hat{\mu}) \mathcal{L}^{(t)}(\alpha) + \frac{1}{N} \sum_{i \in \mathcal{Z}_2} \mathcal{L}^{(t)}(\Xi_i^{(t)}) \right)^2 , \end{aligned} \tag{180}$$

Finally, we use Lemma H.15 and lower bound equation (180) by $\widehat{L}(W^{(t)})^2$. This gives the aimed result.

□

We now present the auxiliary lemmas that helped to establish Lemma H.11. They link the gradient terms with their corresponding loss and are based on Lemma J.20.

Lemma H.12. *Let $t \in [T_1, T]$. Run GD for t iterations. Then, we have:*

$$\sum_{r=1}^{m} \widehat{\ell}^{(t)}(\alpha) \alpha^2 (c_r^{(t)})^2 \geq \tilde{\Omega}(1) \mathcal{L}^{(t)}(\alpha).$$

Proof of Lemma H.12. In order to bound $\sum_{r=1}^{m} \widehat{\ell}^{(t)}(\alpha)\alpha^2(c_r^{(t)})^2$, we apply Lemma J.20. We first verify that the conditions of the lemma are met. From Lemma 4.1 we know that for $t \in [T_0, T]$, we have $c^{(t)} \geq \tilde{\Omega}(1/\alpha)$. Along with Induction hypothesis C.1, this implies that

$$\tilde{\Omega}(1) \leq \tilde{\Omega}(1) - m\tilde{O}(\alpha\sigma_0) \leq \sum_{r=1}^{m} \alpha c_r^{(t)} \leq \tilde{O}(\alpha)m \leq \tilde{O}(1). \tag{181}$$

Therefore, we can apply Lemma J.20 and get the lower bound:

$$\sum_{r=1}^{m} \widehat{\ell}^{(t)}(\alpha)(\alpha c_r^{(t)})^2 \geq \frac{0.05e^{-m\tilde{O}(\sigma_0)}}{\tilde{O}(1)\left(1 + \frac{m^2\tilde{O}(\sigma^2\sigma_0^2 d)}{\tilde{\Omega}(1)^2}\right)} \log\left(1 + e^{-\sum_{r=1}^{m}(\alpha c_r^{(t)})^3}\right) \geq \tilde{\Omega}(1)\mathcal{L}^{(t)}(\alpha). \tag{182}$$

□

Lemma H.13. *Let $t \in [T_1, T]$. Run GD for t iterations. Then, we have:*

$$\frac{1}{N}\sum_{i\in\mathcal{Z}_1}\sum_{r=1}^{m}\sum_{j\neq P(X_i)} \ell_i^{(t)}\left((\alpha^2(c_r^{(t)})^2 + (\Xi_{i,j,r}^{(t)})^2\right) \leq \tilde{O}(1)(1-\hat{\mu})\mathcal{L}^{(t)}(\alpha).$$

Proof of Lemma H.13. We again verify that the conditions of Lemma J.20 are met. By using Induction hypothesis C.1, Induction hypothesis C.2 and Lemma 4.1, we have:

$$\begin{aligned}
&\sum_{r=1}^{m} \alpha c_r^{(t)} + \sum_{r=1}^{m}\sum_{j\neq P(X_i)} y_i \Xi_{i,j,r}^{(t)} \leq m\tilde{O}(\alpha) + mP\tilde{O}(\sigma\sigma_0\sqrt{d}) \leq \tilde{O}(1), \\
&\sum_{r=1}^{m} \alpha c_r^{(t)} + \sum_{r=1}^{m}\sum_{j\neq P(X_i)} y_i \Xi_{i,j,r}^{(t)} \geq \tilde{\Omega}(1) - m\tilde{O}(\alpha\sigma_0) \geq \tilde{\Omega}(1).
\end{aligned} \tag{183}$$

By applying Lemma J.20, we have:

$$\begin{aligned}
&\frac{1}{N}\sum_{i\in\mathcal{Z}_1}\sum_{r=1}^{m}\sum_{j\neq P(X_i)} \ell_i^{(t)}\left((\alpha^2(c_r^{(t)})^2 + (\Xi_{i,j,r}^{(t)})^2\right) \\
&\leq \frac{me^{m\tilde{O}(\sigma_0)}}{\tilde{\Omega}(1)N}\sum_{i\in\mathcal{Z}_1}\log\left(1 + \exp\left(-\sum_{r=1}^{m}\alpha^3(c_r^{(t)})^3 - \Xi_i^{(t)}\right)\right) \\
&\leq \frac{\tilde{O}(1)}{N}\sum_{i\in\mathcal{Z}_1}\log\left(1 + \exp\left(-\sum_{r=1}^{m}\alpha^3(c_r^{(t)})^3 - \Xi_i^{(t)}\right)\right).
\end{aligned} \tag{184}$$

Lastly, we want to link the loss term in equation (184) with $\mathcal{L}^{(t)}(\alpha)$. By applying Induction hypothesis C.1 and Lemma J.24 in equation (184), we finally get:

$$\begin{aligned}
\frac{1}{N}\sum_{i\in\mathcal{Z}_1}\sum_{r=1}^{m}\sum_{j\neq P(X_i)} \ell_i^{(t)}\left((\alpha^2(c_r^{(t)})^2 + (\Xi_{i,j,r}^{(t)})^2\right) &\leq (1-\hat{\mu})(1 + e^{\tilde{O}((\sigma\sigma_0\sqrt{d})^3)})\mathcal{L}^{(t)}(\alpha) \\
&\leq (1-\hat{\mu})\mathcal{L}^{(t)}(\alpha).
\end{aligned} \tag{185}$$

Combining equation (184) and equation (185) yields the aimed result. □

Lemma H.14. *Let $t \in [T_1, T]$. Run GD for t iterations. Then, we have:*

$$\frac{1}{N}\sum_{i\in\mathcal{Z}_2}\sum_{r=1}^{m}\sum_{j\neq P(X_i)} \ell_i^{(t)}(\Xi_{i,j,r}^{(t)})^2 \geq \frac{\tilde{\Omega}(1)}{N}\sum_{i\in\mathcal{Z}_2}\mathcal{L}^{(t)}(\Xi_i^{(t)}).$$

Proof of Lemma H.14. We again verify that the conditions of Lemma J.20 are met. Using Induction hypothesis C.1, Induction hypothesis C.2 and Lemma 4.4, we have:

$$\begin{aligned}
&\sum_{r=1}^{m} \beta c_r^{(t)} + \sum_{r=1}^{m} \sum_{j \neq P(X_i)} y_i \Xi_{i,j,r}^{(t)} \leq m\tilde{O}(\beta) + mP\tilde{O}(1) \leq \tilde{O}(1) \\
&\sum_{r=1}^{m} \beta c_r^{(t)} + \sum_{r=1}^{m} \sum_{j \neq P(X_i)} y_i \Xi_{i,j,r}^{(t)} \geq \tilde{\Omega}(1) - m\tilde{O}(\sigma_0) - mP\tilde{O}(\sigma_0 \sigma\sqrt{d}) \geq \tilde{\Omega}(1).
\end{aligned} \tag{186}$$

By applying Lemma J.20, we have:

$$\begin{aligned}
&\frac{1}{N} \sum_{i \in \mathcal{Z}_2} \sum_{r=1}^{m} \sum_{j \neq P(X_i)} \ell_i^{(t)} (\Xi_{i,j,r}^{(t)})^2 \\
\geq &\frac{0.05 e^{-m\tilde{O}(\sigma\sigma_0\sqrt{d})}}{N\tilde{O}(1)\left(1 + \frac{m^2(\sigma\sigma_0\sqrt{d})^2}{\tilde{\Omega}(1)}\right)} \sum_{i \in \mathcal{Z}_2} \log\left(1 + \exp\left(-\sum_{r=1}^{m} \beta^3 (c_r^{(t)})^3 - \Xi_i^{(t)}\right)\right) \\
\geq &\frac{\tilde{\Omega}(1)}{N} \sum_{i \in \mathcal{Z}_2} \log\left(1 + \exp\left(-\sum_{r=1}^{m} \beta^3 (c_r^{(t)})^3 - \Xi_i^{(t)}\right)\right).
\end{aligned} \tag{187}$$

Lastly, we want to link the loss term in equation (187) with $\mathcal{L}^{(t)}(\Xi_i^{(t)})$. By applying Induction hypothesis C.1 and Lemma J.24 in equation (187), we finally get:

$$\begin{aligned}
\frac{\tilde{\Omega}(1)}{N} \sum_{i \in \mathcal{Z}_2} \sum_{r=1}^{m} \sum_{j \neq P(X_i)} \ell_i^{(t)} (\Xi_{i,j,r}^{(t)})^2 &\geq \frac{\tilde{\Omega}(1) e^{-m\tilde{O}(\beta^3)}}{N} \sum_{i \in \mathcal{Z}_2} \mathcal{L}^{(t)}(\Xi_i^{(t)}) \\
&\geq \frac{\tilde{\Omega}(1)}{N} \sum_{i \in \mathcal{Z}_2} \mathcal{L}^{(t)}(\Xi_i^{(t)}).
\end{aligned} \tag{188}$$

Combining equation (187) and equation (188) yields the aimed result. □

Lemma H.15. *Let $t \in [0, T]$ Run GD for for t iterations. Then, we have:*

$$(1 - \hat{\mu})\mathcal{L}^{(t)}(\alpha) + \frac{1}{N} \sum_{i \in \mathcal{Z}_2} \mathcal{L}^{(t)}(\Xi_i^{(t)}) \geq \Theta(1) \widehat{L}(W^{(t)}). \tag{189}$$

Proof of Lemma H.15. we need to lower bound $\mathcal{L}^{(t)}(\alpha)$. By successively applying Lemma J.24 and Induction hypothesis C.1, we obtain:

$$\begin{aligned}
(1 - \hat{\mu})\mathcal{L}^{(t)}(\alpha) &= \frac{1}{N} \sum_{i \in \mathcal{Z}_1} \frac{1 + e^{-\Xi_i^{(t)}}}{1 + e^{-\Xi_i^{(t)}}} \log\left(1 + \exp\left(-\sum_{r=1}^{m} (\alpha c_r^{(t)})^3\right)\right) \\
&\geq \frac{1}{N} \sum_{i \in \mathcal{Z}_1} \frac{1}{1 + e^{-\Xi_i^{(t)}}} \log\left(1 + \exp\left(-\sum_{r=1}^{m} (\alpha c_r^{(t)})^3\right) - \Xi_i^{(t)}\right) \\
&\geq \frac{\widehat{L}_{\mathcal{Z}_1}(W^{(t)})}{1 + e^{\tilde{O}((\sigma\sigma_0\sqrt{d})^3)}} \\
&\geq \Theta(1) \widehat{L}_{\mathcal{Z}_1}(W^{(t)}).
\end{aligned} \tag{190}$$

By successively applying Lemma J.24 and Induction hypothesis C.1, we obtain:

$$\begin{aligned}
\frac{1}{N}\sum_{i\in\mathcal{Z}_2}\mathcal{L}^{(t)}(\Xi_i^{(t)}) &= \frac{1}{N}\sum_{i\in\mathcal{Z}_2}\frac{1+e^{-\sum_{r=1}^m(\beta c_r^{(t)})^3}}{1+e^{-\sum_{r=1}^m(\beta c_r^{(t)})^3}}\log\left(1+\exp\left(-\Xi_i^{(t)}\right)\right)\\
&\geq \frac{1}{N}\sum_{i\in\mathcal{Z}_2}\frac{1}{1+e^{-\sum_{r=1}^m(\beta c_r^{(t)})^3}}\log\left(1+\exp\left(-\sum_{r=1}^m(\beta c_r^{(t)})^3\right)-\Xi_i^{(t)}\right)\\
&\geq \frac{\widehat{L}_{\mathcal{Z}_2}(W^{(t)})}{1+e^{\tilde{O}((\beta\sigma_0)^3)}}\\
&\geq \Theta(1)\widehat{L}_{\mathcal{Z}_2}(W^{(t)}). \qquad (191)
\end{aligned}$$

Combining equation (190) and equation (191) yields the aimed result.

□

I AUXILIARY LEMMAS FOR GD+M

This section presents the auxiliary lemmas needed in Appendix F. These lemmas mainly consists in rewriting of the GD update on the signal and noise components and consequences of such rewriting.

I.1 PROJECTION ONTO THE SIGNAL COMPONENT

Lemma I.1 (Bound on derivative for GD+M)**.** *Let $i \in \mathcal{Z}$. Then, $\ell_i^{(t)} = \Theta(1)\widehat{\ell}^{(t)}(\theta)$.*

Proof. Let $i \in [N]$. Using Induction hypothesis C.4, we have:

$$\ell_i^{(t)} = \text{sigmoid}\left(-\theta^3 \sum_{s=1}^{m} (c_s^{(t)})^3 - \sum_{s=1}^{m} \sum_{j \neq P(X_i)} (\Xi_{i,j,s}^{(t)})^3\right).$$

Therefore, we deduce that:

$$e^{-\tilde{O}((\sigma\sigma_0\sqrt{d})^3)}\widehat{\ell}^{(t)}(\theta) \leq \ell_i^{(t)} \leq e^{\tilde{O}((\sigma\sigma_0\sqrt{d})^3)}\widehat{\ell}^{(t)}(\theta)$$

which yields the aimed result.

□

I.1.1 GRADIENT TERMS

Lemma I.2. *Using GD+M, the signal sequence $c_r^{(t)}$ is non-decreasing for all $r \in [m]$.*

Proof of Lemma I.2. From Lemma F.1, we know that the signal momentum is equal to:

$$-\mathcal{G}_r^{(t)} = \tilde{\Theta}(1)(1-\gamma)\sum_{\tau=0}^{t-1}\left((1-\hat{\mu})\alpha^3\widehat{\ell}^{(\tau)}(\alpha) + \hat{\mu}\beta^3\widehat{\ell}^{(\tau)}(\beta)\right)\gamma^{t-1-\tau}(c_r^{(\tau)})^2 \geq 0. \tag{192}$$

Since $c_r^{(t+1)} - c_r^{(t)} = -\eta\mathcal{G}_r^{(t)}$, we proved that the increment is non-negative.

□

I.2 PROJECTION ONTO THE NOISE COMPONENT

Lemma I.3 (Bound on noise momentum)**.** *Run GD+M on the loss function $\widehat{L}(W)$. Let $i \in [N]$, $j \in [P]\backslash\{P(X_i)\}$. At a time t, the noise momentum is bounded with probability $1 - o(1)$ as:*

$$\left|-G_{i,j,r}^{(t+1)} + \gamma G_{i,j,r}^{(t)}\right| \leq (1-\gamma)\tilde{O}(\sigma^4\sigma_0^2 d^2)\nu^{(t)}.$$

Proof of Lemma I.3. Let $i \in [N]$ and $j \in [P]\backslash\{P(X_i)\}$. Combining the equation (GDM-N) update rule and Lemma D.3 to get the noise gradient $\mathsf{G}_{i,j,r}^{(t)}$, we obtain

$$\begin{aligned}&\left|-G_{i,j,r}^{(t+1)} + \gamma G_{i,j,r}^{(t)}\right| \\ &\leq \frac{3(1-\gamma)}{N}\ell_i^{(t)}(\Xi_{i,j,r}^{(t)})^2\|X_j^{(i)}\|_2^2 + \left|\frac{3(1-\gamma)}{N}\sum_{a=1}^{N}\ell_a^{(t)}\sum_{k\neq P(X_a)}(\Xi_{a,k,r}^{(t)})^2\langle X_a[k], X_i[j]\rangle\right|.\end{aligned} \tag{193}$$

Using Lemma J.5 and Lemma J.7, equation (193) becomes with probability $1 - o(1)$,

$$\begin{aligned}&\left|-G_{i,j,r}^{(t+1)} + \gamma G_{i,j,r}^{(t)}\right| \\ &\leq \frac{(1-\gamma)\tilde{\Theta}(\sigma^2 d)}{N}\ell_i^{(t)}(\Xi_{i,j,r}^{(t)})^2 + \frac{(1-\gamma)\tilde{\Theta}(\sigma^2\sqrt{d})}{N}\sum_{a=1}^{N}\ell_a^{(t)}\sum_{k\neq P(X_a)}(\Xi_{a,k,r}^{(t)})^2\end{aligned} \tag{194}$$

Using $\ell_i^{(t)}/N \leq \nu^{(t)}$, Induction hypothesis C.4, we upper bound the first term in equation (194) to get:

$$\begin{aligned}&\left|-G_{i,j,r}^{(t+1)} + \gamma G_{i,j,r}^{(t)}\right| \\ &\leq (1-\gamma)\tilde{O}(\sigma^4\sigma_0^2 d^2)\nu^{(t)} + \frac{(1-\gamma)\tilde{\Theta}(\sigma^2\sqrt{d})}{N}\sum_{a=1}^{N}\ell_a^{(t)}\sum_{k\neq P(X_a)}(\Xi_{a,k,r}^{(t)})^2.\end{aligned} \tag{195}$$

We upper bound the second term in equation (195) by again using Induction hypothesis C.4:

$$\left|-G_{i,j,r}^{(t+1)} + \gamma G_{i,j,r}^{(t)}\right| \leq (1-\gamma)\left(\tilde{O}(\sigma^4\sigma_0^2 d^2) + \tilde{O}(P\sigma_0^2\sigma^4 d^{3/2})\right)\nu^{(t)} \tag{196}$$

By using $P \leq \tilde{O}(1)$ and thus, $\tilde{O}(P\sigma_0^2\sigma^4 d^{3/2}) \leq \tilde{O}(\sigma^4\sigma_0^2 d^2)$ in equation (196), we obtain the desired result. □

Lemma I.4. *For all $t, t' \in [T]$ such that $t' \leq t$ the noise momentum is bounded as*

$$|G_{i,j,r}^{(t)}| \leq |G_{i,j,r}^{(t')}| + (1-\gamma)\tilde{O}(\sigma^4\sigma_0^2 d^2)\sum_{\tau=t'}^{t-1}\gamma^{t-1-\tau}\nu^{(\tau)}.$$

Proof of Lemma I.4. Let $\tau \in [T]$. From Lemma I.3, we know that:

$$|G_{i,j,r}^{(\tau+1)}| \leq |\gamma G_{i,j,r}^{(\tau)}| + (1-\gamma)\tilde{O}(\sigma^4\sigma_0^2 d^2)\nu^{(\tau)}. \tag{197}$$

We unravel the recursion equation (197) rule for $\tau = t', \ldots, t-1$ and obtain:

$$|G_{i,j,r}^{(t)}| \leq |G_{i,j,r}^{(t')}| + (1-\gamma)\tilde{O}(\sigma^4\sigma_0^2 d^2)\sum_{\tau=t'}^{t-1}\gamma^{t-1-\tau}\nu^{(\tau)}. \tag{198}$$

□

I.3 Convergence rate of the training loss using GD+M

In this section, we prove that when using GD+M, the training loss converges sublinearly in our setting.

Lemma I.5 (Convergence rate of the loss)**.** *For $t \in [\mathcal{T}_1, T]$ Using GD+M with learning rate $\eta \in (0, 1/L)$, the loss sublinearly converges to zero as*

$$(1-\hat{\mu})\widehat{\mathcal{L}}^{(t)}(\alpha) + \hat{\mu}\widehat{\mathcal{L}}^{(t)}(\beta) \leq \tilde{O}\left(\frac{1}{\eta\beta^2(t-\mathcal{T}_1+1)}\right). \tag{199}$$

Proof of Lemma I.5. Let $t \in [\mathcal{T}_1, T]$. From Lemma I.6, we know that the signal gradient is bounded as $-\mathscr{G}^{(t)} \geq -\mathscr{G}^{(s)}$ for $s \in [\mathcal{T}_1, t]$.

$$\begin{aligned}-\mathcal{G}^{(t)} &= -\gamma^{t-\mathcal{T}_1}\mathcal{G}^{(\mathcal{T}_1)} - (1-\gamma)\sum_{s=\mathcal{T}_1}^{t}\gamma^{t-s}\mathscr{G}^{(s)} \\ &\geq -(1-\gamma)\sum_{s=\mathcal{T}_1}^{t}\gamma^{t-s}\mathscr{G}^{(s)} \\ &\geq -(1-\gamma)\mathscr{G}^{(t)}\sum_{s=\mathcal{T}_1}^{t}\gamma^{t-s} \\ &= -\Theta(1)\mathscr{G}^{(t)}.\end{aligned} \tag{200}$$

From Lemma D.2, the signal gradient is:

$$-\mathscr{G}^{(t)} = \Theta(1)\left(\alpha^3\widehat{\ell}^{(t)}(\alpha) + \beta^3\widehat{\ell}^{(t)}(\beta)\right)(c^{(t)})^2. \tag{201}$$

From Lemma 5.3, we know that $c^{(t)} \geq \tilde{\Omega}(1/\beta)$. Thus, we simplify equation (201) as:

$$-\mathscr{G}^{(t)} \geq \tilde{\Omega}(\beta) \left((1-\hat{\mu}) \widehat{\ell}^{(t)}(\alpha) + \hat{\mu}\beta \widehat{\ell}^{(t)}(\beta) \right). \tag{202}$$

By combining equation (200) and equation (202), we finally obtain:

$$-\mathcal{G}^{(t)} \geq \tilde{\Omega}(\beta) \left((1-\hat{\mu}) \widehat{\ell}^{(t)}(\alpha) + \hat{\mu} \widehat{\ell}^{(t)}(\beta) \right). \tag{203}$$

We now plug equation (203) in the signal update equation (GDM-S).

$$c^{(t+1)} \geq c^{(t)} + \tilde{\Omega}(\eta\beta) \left((1-\hat{\mu}) \widehat{\ell}^{(t)}(\alpha) + \hat{\mu} \widehat{\ell}^{(t)}(\beta) \right). \tag{204}$$

We now apply Lemma J.22 to lower bound equation (204) by loss terms. We have:

$$c^{(t+1)} \geq c^{(t)} + \tilde{\Omega}(\eta\beta) \left((1-\hat{\mu}) \widehat{\mathcal{L}}^{(t)}(\alpha) + \hat{\mu} \widehat{\mathcal{L}}^{(t)}(\beta) \right). \tag{205}$$

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

$$(1-\hat{\mu}) \widehat{\mathcal{L}}^{(t)}(\alpha) + \hat{\mu} \widehat{\mathcal{L}}^{(t)}(\beta) > \frac{\tilde{O}(1)}{\eta\beta^2(t-\mathcal{T}_1+1)}. \tag{206}$$

From the equation (GDM-S) update, we know that $c_r^{(\tau)}$ is a non-decreasing sequence which implies that $\sum_{r=1}^m (\theta c_r^{(\tau)})^3$ is also non-decreasing for $\tau \in [T]$. Since $x \mapsto \log(1+\exp(-x))$ is non-increasing, this implies that for $s \leq t$, we have:

$$\frac{\tilde{O}(1)}{\eta\beta^2(t-\mathcal{T}_1)} < (1-\hat{\mu}) \widehat{\mathcal{L}}^{(t)}(\alpha) + \hat{\mu} \widehat{\mathcal{L}}^{(t)}(\beta) \leq (1-\hat{\mu}) \widehat{\mathcal{L}}^{(s)}(\alpha) + \hat{\mu} \widehat{\mathcal{L}}^{(s)}(\beta). \tag{207}$$

Plugging equation (207) in the update equation (205) yields for $s \in [\mathcal{T}_1, t]$:

$$c^{(s+1)} > c^{(s)} + \frac{\tilde{O}(1)}{\beta(t-\mathcal{T}_1+1)} \tag{208}$$

We now sum equation (208) for $s = \mathcal{T}_1, \dots, t-1$ and obtain:

$$c^{(t)} > c^{(\mathcal{T}_1)} + \frac{\tilde{O}(1)(t-\mathcal{T}_1)}{\beta(t-\mathcal{T}_1+1)} > \frac{\tilde{O}(1)}{\beta}, \tag{209}$$

where we used the fact that $c^{(\mathcal{T}_1)} \geq \tilde{\Omega}(1/\beta) \geq 0$ (Lemma 5.2) in the last inequality. Let's now show that equation (209) implies that equation (206) is a contradiction. Indeed, we have:

$$\begin{aligned}
&\eta\beta^2(t-\mathcal{T}_1+1) \left((1-\hat{\mu}) \widehat{\mathcal{L}}^{(t)}(\alpha) + \hat{\mu} \widehat{\mathcal{L}}^{(t)}(\beta) \right) \\
\leq &\eta\beta^2 T \left((1-\hat{\mu}) \log \left(1 + \exp(-(\alpha c^{(t)})^3 - \sum_{r \neq r_{\max}} (\alpha c_r^{(t)})^3 \right) \right. \\
&\left. + \hat{\mu} \log \left(1 + \exp(-(\beta c^{(t)})^3 - \sum_{r \neq r_{\max}} (\beta c_r^{(t)})^3 \right) \right) \\
\leq &\eta\beta^2 T \left((1-\hat{\mu}) \log \left(1 + \exp(-\tilde{O}(\alpha^3/\beta^3) \right) + \hat{\mu} \log \left(1 + \exp(-\tilde{O}(1) \right) \right),
\end{aligned} \tag{210}$$

where we used $\sum_{r \neq r_{\max}} (c_r^{(t)})^3 \geq -m\tilde{O}(\sigma_0^3)$ along with equation (209) in equation (210). We now apply Lemma J.22 in equation (210) and obtain:

$$\eta\beta^2(t-\mathcal{T}_1+1) \left((1-\hat{\mu}) \widehat{\mathcal{L}}^{(t)}(\alpha) + \hat{\mu} \widehat{\mathcal{L}}^{(t)}(\beta) \right) \leq \frac{(1-\hat{\mu})\eta\beta^2 T}{1+\exp(\tilde{O}(\alpha^3/\beta^3))} + \frac{\hat{\mu}\eta\beta^2 T}{1+\exp(\tilde{O}(1))}. \tag{211}$$

Since $T \leq d^{O(\log d)}/\eta$, $\hat{\mu} = \Theta(1/N)$, $\beta = \tilde{\Theta}(1/\sqrt{d})$, $\alpha = \tilde{\Theta}(d)$, we finally have:

$$\begin{aligned}
&\eta\beta^2(t-\mathcal{T}_1+1) \left((1-\hat{\mu}) \widehat{\mathcal{L}}^{(t)}(\alpha) + \hat{\mu} \widehat{\mathcal{L}}^{(t)}(\beta) \right) \\
\leq &\frac{\exp(O(\log^2 d))}{d(1+\exp(\tilde{O}(d^{9/2}))} + \frac{\exp(O(\log^2 d))}{\mathrm{poly}(d)(1+\exp(O(\mathrm{polylog}(d))))} \\
\leq &o(1) < \tilde{O}(1),
\end{aligned} \tag{212}$$

which contradicts equation (206). □

We now provide an auxiliary lemma needed to obtain equation (I.5).

Lemma I.6. *Let $t \in [\mathcal{T}_1, T]$. Then, the signal gradient decreases i.e. $-\mathscr{G}^{(s)} \geq -\mathscr{G}^{(t)}$ for $s \in [\mathcal{T}_1, t]$.*

Proof of Lemma I.6. From Lemma D.2, we know that

$$-\mathscr{G}^{(t)} = \Theta(1)\left(\alpha^3 \widehat{\ell}^{(t)}(\alpha) + \beta^3 \widehat{\ell}^{(t)}(\beta)\right)(c^{(t)})^2. \tag{213}$$

Since $c_r^{(t)} \geq -\tilde{O}(\sigma_0)$, we bound equation (213) as:

$$-\mathscr{G}^{(t)} \leq \Theta(1)\left(\alpha^3 \mathfrak{S}((\alpha c^{(t)})^3) + \beta^3 \mathfrak{S}((\beta c^{(t)})^3)\right)(c^{(t)})^2. \tag{214}$$

The function $x \mapsto x^2\mathfrak{S}(x^3)$ is non-increasing for $x \geq 1$. Since $c^{(t)} \geq \tilde{\Omega}(1/\beta)$, we have:

$$-\mathscr{G}^{(t)} \leq \Theta(1)\left(\alpha^3 \mathfrak{S}((\alpha c^{(\mathcal{T}_1)})^3) + \beta^3 \mathfrak{S}((\beta c^{(\mathcal{T}_1)})^3)\right)(c^{(\mathcal{T}_1)})^2 = -\mathscr{G}^{(\mathcal{T}_1)}. \tag{215}$$

□

J USEFUL LEMMAS

J.1 PROBABILISTIC LEMMAS

In this section, we introduce the probabilistic lemmas used in the proof. In subsubsection J.1.1, we introduce some high-probability bounds and properties of sub-Gaussian and sub-exponential random variables. subsubsection J.1.2 reminds the anti-concentration property of Gaussian polynomials. Lastly, subsubsection J.1.3 presents some properties satisfied by the cube of a Gaussian random variable.

J.1.1 HIGH-PROBABILITY BOUNDS

Lemma J.1. *The sum of of symmetric random variables is symmetric.*

Lemma J.2 (Sum of sub-Gaussians (Vershynin, 2018))**.** *Let $\sigma_1, \sigma_2 > 0$. Let X and Y respetively be σ_1- and σ_2-subGaussian random variables. Then, $X + Y$ is $\sqrt{\sigma_1 + \sigma_2}$-subGaussian random variable.*

Lemma J.3 (High probability bound subGaussian (Vershynin, 2018))**.** *Let $t > 0$. Let X be a σ-subGaussian random variable. Then, we have:*

$$\mathbb{P}\left[|X| > t\right] \leq 2e^{-\frac{t^2}{2\sigma^2}}.$$

Theorem J.1 (Concentration of Lipschitz functions of Gaussian variables (Wainwright, 2019))**.** *Let $X_1, \ldots, X_N$ be N i.i.d. random variables such that $X_i \sim \mathcal{N}(0, \sigma^2)$ and $X := (X_1, \ldots, X_n)$. Let $f \colon \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz with respect to the Euclidean norm. Then,*

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{t^2}{2L}}. \tag{216}$$

Lemma J.4 (Expectation of Gaussian vector (Wainwright, 2019))**.** *Let $X \in \mathbb{R}^d$ be a Gaussian vector such that $X \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. Then, its expectation is equal to $\mathbb{E}[\|X\|_2] = \Theta(\sigma\sqrt{d})$.*

Lemma J.5 (High-probability bound on squared norm of Gaussian)**.** *Let $X \in \mathbb{R}^d$ be a Gaussian vector such that $X \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. Then, with probability at least $1 - o(1)$, we have $\|X\|_2^2 = \Theta(\sigma^2 d)$.*

Proof of Lemma J.5. We know that the $\|\cdot\|_2$ is 1-Lipschitz and by applying Theorem J.1, we therefore have::

$$\mathbb{P}\left[|\|X\|_2 - \mathbb{E}[\|X\|_2]| > \epsilon\right] \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right). \tag{217}$$

By rewriting equation (217) and using Lemma J.4, we have with probability $1 - \delta$,

$$\Theta(\sigma\sqrt{d}) - \sigma\sqrt{2\log\left(\frac{1}{\delta}\right)} \leq \|X\|_2 \leq \Theta(\sigma\sqrt{d}) + \sigma\sqrt{2\log\left(\frac{1}{\delta}\right)}. \tag{218}$$

By squaring equation (218) and using $(a+b)^2 \leq a^2 + b^2$, we obtain the aimed result. □

Lemma J.6 (Precise bound on squared norm of Gaussian)**.** *Let $X \in \mathbb{R}^d$ be a Gaussian vector such that $X \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$. Then, we have:*

$$\mathbb{P}\left[\|X\|_2 \in \left[\frac{1}{2}\sigma\sqrt{d}, \frac{3}{2}\sigma\sqrt{d}\right]\right] \geq 1 - e^{-d/8}.$$

Proof of Lemma J.6. We know that the $\|\cdot\|_2$ is 1-Lipschitz and by applying Theorem J.1, we therefore have:

$$\mathbb{P}\left[|\|X\|_2 - \mathbb{E}[\|X\|_2]| > \epsilon\right] \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right). \tag{219}$$

We use Lemma J.4 and set $\epsilon = \frac{\sigma\sqrt{d}}{2}$ in equation (219) to finally get:

$$\mathbb{P}\left[|\|X\|_2 - \mathbb{E}[\|X\|_2]| > \frac{\sigma\sqrt{d}}{2}\right] \leq \exp\left(-\frac{d}{8}\right).$$

□

Lemma J.7 (High-probability bound on dot-product of Gaussians)**.** *Let X and Y be two independent Gaussian vectors in $\mathbb{R}^d$ such that X, Y independent and $X \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ and $Y \sim \mathcal{N}(0, \sigma_0^2\mathbf{I})$. Assume that $\sigma\sigma_0 \leq 1/d$. Then, with probability $1 - o(1)$, we have:*

$$|\langle X, Y\rangle| \leq \tilde{O}(\sigma\sigma_0\sqrt{d}).$$

Proof of Lemma J.7 . Let's define $Z := \langle X, Y\rangle$. We first remark that Z is a sub-exponential random variable. Indeed, the generating moment function is:

$$M_Z(t) = \mathbb{E}[e^{t\langle X,Y\rangle}] = \frac{1}{(1-\sigma^2\sigma_0^2t^2)^{d/2}} = e^{-\frac{d}{2}\log(1-\sigma^2\sigma_0^2t^2)} \leq e^{\frac{d\sigma^2\sigma_0^2t^2}{2}}, \qquad \text{for } t \leq \frac{1}{\sigma\sigma_0}.$$

where we used $\log(1-x) \geq -x$ for $x < 1$ in the last inequality. Therefore, by definition of a sub-exponential variable, we have:

$$\mathbb{P}\left[|Z - \mathbb{E}[Z]| > \epsilon\right] \leq \begin{cases} 2e^{-\frac{\epsilon^2}{2d\sigma^2\sigma_0^2}} & \text{for } 0 \leq \epsilon \leq d\sigma\sigma_0 \\ 2e^{-\frac{\epsilon}{2\sigma\sigma_0}} & \text{for } \epsilon \geq d\sigma\sigma_0 \end{cases}. \tag{220}$$

Since $\sigma^2 d \leq 1$ and $\epsilon \in [0, 1]$, equation (220) is bounded as:

$$\mathbb{P}\left[|Z - \mathbb{E}[Z]| > \epsilon\right] \leq 2e^{-\frac{\epsilon^2}{2d\sigma^2\sigma_0^2}}. \tag{221}$$

We know that $\mathbb{E}[Z] = M'(0) = \left(d(1-\sigma^2\sigma_0^2t^2)^{-\frac{d}{2}-1}\sigma^2\sigma_0^2t\right)(0) = 0$. By plugging this expectation in equation (221), we have with probability $1 - \delta$,

$$|\langle X, Y\rangle| \leq \sigma\sigma_0\sqrt{2d\log\left(\frac{2}{\delta}\right)}.$$

□

Lemma J.8 (High-probability bound on dot-product of Gaussians)**.** *Let X and Y be two independent Gaussian vectors in $\mathbb{R}^d$ such that $X, Y \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. Then, with probability $1 - \delta$, we have:*

$$\left|\left\langle \frac{X}{\|X\|_2}, Y\right\rangle\right| \leq \tilde{O}(\sigma).$$

Proof of Lemma J.7 . Let $U := X/\|X\|_2$ and $Z := \langle U, Y\rangle$. We know that the pdf of U in polar coordinates is $f_U(\theta) = \frac{\Gamma(d/2)}{2\pi^{d/2}}$. Therefore, the generating moment function of Z is:

$$\begin{aligned}
M_Z(t) &= \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}^d} e^{t\langle u,y\rangle} f_U(u) f_Y(y) dudy \\
&= \frac{\Gamma(d/2)}{2\pi^{d/2}(2\pi\sigma^2)^{d/2}} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}^d} e^{t\langle u,y\rangle} e^{-\frac{\|y\|_2^2}{2\sigma^2}} dydu \\
&= \frac{\Gamma(d/2)}{2\pi^{d/2}(2\pi\sigma^2)^{d/2}} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}^d} e^{-\frac{\|y - t\sigma^2 u\|_2^2}{2\sigma^2}} e^{\frac{t^2\sigma^2\|u\|_2^2}{2}} dydu \\
&= \frac{\Gamma(d/2)}{2\pi^{d/2}(2\pi\sigma^2)^{d/2}} \int_{\mathbb{S}^{d-1}} e^{\frac{\sigma^2 t^2\|u\|_2^2}{2}} du \\
&= \frac{\Gamma(d/2)}{2\pi^{d/2}(2\pi\sigma^2)^{d/2}} \int_{\mathbb{S}^{d-1}} e^{\frac{\sigma^2 t^2}{2}} du \\
&= e^{\frac{\sigma^2 t^2}{2}}.
\end{aligned} \tag{222}$$

equation (222) indicates that Z is a sub-Gaussian random variable of parameter σ. By definition, it satisfies

$$\mathbb{P}[|Z| > \epsilon] \leq 2e^{-\frac{\epsilon^2}{2\sigma^2}}. \tag{223}$$

Setting $\delta = 2e^{-\frac{\epsilon^2}{2\sigma^2}}$ in equation (223) yields that we have with probability $1-\delta$,

$$\left|\left\langle \frac{X}{\|X\|_2}, Y \right\rangle\right| \leq \sqrt{2\log\left(\frac{2}{\delta}\right)}.$$

□

Lemma J.9 (High probability bound for ratio of norms)**.** *Let $X_1, \ldots, X_n$ i.i.d. vectors from $\mathcal{N}(0, \sigma^2\mathbf{I})$. Then, from Lemma J.6, we have:*

$$\frac{\|X_1\|_2^2}{\|\sum_{i=1}^n X_i\|_2} = \tilde{\Theta}\left(\sigma\sqrt{\frac{d}{n}}\right). \tag{224}$$

Proof of Lemma J.9. We know that for $X_1 \sim \mathcal{N}(0, \sigma^2 d)$, we have:

$$\mathbb{P}\left[\|X_1\|_2^2 \in \left[\frac{\sigma^2 d}{4}, \frac{9\sigma^2 d}{4}\right]\right] \leq e^{-d/8}. \tag{225}$$

Therefore, using the law of total probability and equation (225), we have:

$$\begin{aligned}
\mathbb{P}\left[\frac{\|X_1\|_2^2}{\|\sum_{i=1}^n X_i\|_2} > t\right] &= \mathbb{P}\left[\frac{\|X_1\|_2^2}{\|\sum_{i=1}^n X_i\|_2} > t \,\middle|\, \|X_1\|_2^2 > \frac{9\sigma^2 d}{4}\right] \mathbb{P}\left[\|X_1\|_2^2 > \frac{9\sigma^2 d}{4}\right] \\
&\quad + \mathbb{P}\left[\frac{\|X_1\|_2^2}{\|\sum_{i=1}^n X_i\|_2} > t \,\middle|\, \|X_1\|_2^2 < \frac{9\sigma^2 d}{4}\right] \mathbb{P}\left[\|X_1\|_2^2 < \frac{9\sigma^2 d}{4}\right] \\
&\leq e^{-d/8} + \mathbb{P}\left[\frac{\|X_1\|_2^2}{\|\sum_{i=1}^n X_i\|_2} > t \,\middle|\, \|X_1\|_2^2 < \frac{9\sigma^2 d}{4}\right].
\end{aligned} \tag{226}$$

Now, we can further bound equation (226) as:

$$\mathbb{P}\left[\frac{\|X_1\|_2^2}{\|\sum_{i=1}^n X_i\|_2} > t\right] \leq e^{-d/8} + \mathbb{P}\left[\frac{9\sigma^2 d}{4t} > \|\sum_{i=1}^n X_i\|_2\right]. \tag{227}$$

Since $\sum_{i=1}^n X_i \sim \mathcal{N}(0, n\sigma^2)$, we also have

$$\mathbb{P}\left[\|\sum_{i=1}^n X_i\|_2 \in \left[\frac{\sigma\sqrt{nd}}{2}, \frac{3\sigma\sqrt{nd}}{2}\right]\right] \leq e^{-d/8}. \tag{228}$$

Therefore by setting $t = \frac{3\sigma}{2}\sqrt{\frac{d}{n}}$, we obtain:

$$\mathbb{P}\left[\frac{\|X_1\|_2^2}{\|\sum_{i=1}^n X_i\|_2} > \frac{3\sigma}{2}\sqrt{\frac{d}{n}}\right] \leq 2e^{-d/8}. \tag{229}$$

Doing the similar reasoning for the lower bound yields:

$$\mathbb{P}\left[\frac{\|X_1\|_2^2}{\|\sum_{i=1}^n X_i\|_2} < \frac{\sigma}{2}\sqrt{\frac{d}{n}}\right] \leq 2e^{-d/8}. \tag{230}$$

□

Lemma J.10 (High probability bound norms vs dot product)**.** *Let $X_1, \ldots, X_n$ i.i.d. vectors from $\mathcal{N}(0, \sigma^2\mathbf{I})$. Then, with probability $1 - o(1)$, we have:*

$$\frac{\sqrt{d}|\langle X_1, X_2\rangle|}{\|\sum_{i=1}^N X_i\|_2} \leq \frac{\|X_1\|_2^2}{\|\sum_{i=1}^N X_i\|_2}. \tag{231}$$

Proof of Lemma J.10. To show the result, it's enough to upper bound the following probability:

$$\mathbb{P}\left[\|X_1\|_2^2 > \sqrt{d}|\langle X_1, X_2\rangle|\right]. \tag{232}$$

By using the law of total probability we have:

$$\begin{aligned}
&\mathbb{P}\left[\|X_1\|_2^2 > \sqrt{d}|\langle X_1, X_2\rangle|\right] \\
=&\mathbb{P}\left[\|X_1\|_2^2 > \sqrt{d}|\langle X_1, X_2\rangle| \;\middle|\; \|X_1\|_2^2 \in \left[\frac{\sigma^2 d}{2}, \frac{9\sigma^2 d}{4}\right]\right]\mathbb{P}\left[\|X_1\|_2^2 \in \left[\frac{\sigma^2 d}{2}, \frac{9\sigma^2}{4}\right]\right] \\
+&\mathbb{P}\left[\|X_1\|_2^2 > \sqrt{d}|\langle X_1, X_2\rangle| \;\middle|\; \|X_1\|_2^2 \notin \left[\frac{\sigma^2 d}{2}, \frac{9\sigma^2 d}{4}\right]\right]\mathbb{P}\left[\|X_1\|_2^2 \notin \left[\frac{\sigma^2 d}{2}, \frac{9\sigma^2}{4}\right]\right] \\
\leq&\mathbb{P}\left[\|X_1\|_2^2 > \sqrt{d}|\langle X_1, X_2\rangle| \;\middle|\; \|X_1\|_2^2 \in \left[\frac{\sigma^2 d}{2}, \frac{9\sigma^2 d}{4}\right]\right] + e^{-d/8},
\end{aligned} \tag{233}$$

where we used Lemma J.6 in equation (233). Using Lemma J.6 again, we can simplify equation (233) as:

$$\begin{aligned}
\mathbb{P}\left[\|X_1\|_2^2 > \sqrt{d}|\langle X_1, X_2\rangle|\right] &\leq \mathbb{P}\left[\frac{9\sigma^2\sqrt{d}}{4} > |\langle X_1, X_2\rangle|\right] + e^{-d/8} \\
&\leq 2e^{-d/8}.
\end{aligned}$$

□

J.1.2 Anti-concentration of Gaussian polynomials

Theorem J.2 (Anti-concentration of Gaussian polynomials (Carbery & Wright, 2001; Lovett, 2010))**.** *Let $P(x) = P(x_1, \ldots, x_n)$ be a degree d polynomial and $x_1, \ldots, x_n$ be i.i.d. Gaussian univariate random variables. Then, the following holds for all d, n.*

$$\mathbb{P}\left[|P(x)| \leq \epsilon \mathrm{Var}[P(x)]^{1/2}\right] \leq O(d)\epsilon^{1/d}.$$

Lemma J.11 (Gaussians and Hermite (Lovett, 2010))**.** *Let $\mathcal{P}(x_1, \ldots, x_P) = \sum_{k=1}^d \sum_{\mathcal{I}\subset[P]:|\mathcal{I}|=k} c_{\mathcal{I}} \prod_{i\in\mathcal{I}} x_i$ be a degree d polynomial where $x_1, \ldots, x_P \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2\mathbf{I})$ and $c_{\mathcal{I}} \in \mathbb{R}$.*
Let $\mathcal{H}(x) = \sum_{e\in\mathbb{N}^P:|e|\leq d} c_e^H \prod_{i=1}^P H_{e_i}(x_i)$ be the corresponding Hermite polynomial to $\mathcal{P}$ where $\{H_{e_k}\}_{k=1}^d$ is the Hermite polynomial basis. Then, the variance of P is given by $\mathrm{Var}[P(x)^2] = \sum_e |c_e^H|^2$.

Lemma J.12. *Let $\{v_r\}_{r=1}^m$ be vectors in $\mathbb{R}^d$ such that there exist a unit norm vector x that satisfies $|\sum_{r=1}^m \langle v_r, x\rangle^3| \geq 1$. Then, for $\xi_1, \dots, \xi_k \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ i.i.d., we have:*

$$\mathbb{P}\left[\left|\sum_{j=1}^P \sum_{r=1}^m \langle v_r, \xi_j\rangle^3\right| \geq \tilde{\Omega}(\sigma^3)\right] \geq 1 - \frac{O(d)}{2^{1/d}}.$$

Proof of Lemma J.12. Let $\xi_1, \dots, \xi_j \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ i.i.d. We decompose ξ_j as $\xi_j = \tilde{a}_j x + b_j$ where b_j is an independent Gaussian on the orthogonal complement of x and $\tilde{a}_j \sim \mathcal{N}(0, \sigma^2)$. Finally, we rewrite $\tilde{a}_j$ as $\tilde{a}_j = \sigma a_j$ where $a_j \sim \mathcal{N}(0, 1)$. Therefore, we can rewrite $\sum_{j=1}^P \sum_{r=1}^m \langle v_r, \xi_j\rangle^3$ as a polynomial $\mathcal{P}(a_1, \dots, a_P)$ defined as:

$$\begin{aligned}\mathcal{P}(a_1, \dots, a_P) = \sigma^3 \sum_{j=1}^P a_j^3 \left(\sum_{r=1}^m \langle v_r, x\rangle^3\right) + 3\sigma^2 \sum_{j=1}^P a_j^2 \left(\sum_{r=1}^m \langle v_r, x\rangle^2 \langle v_r, b_j\rangle\right) \\ + 3\sigma \sum_{j=1}^P a_j \left(\sum_{r=1}^m \langle v_r, x\rangle \langle v_r, b_j\rangle^2\right) + \sum_{j=1}^P \sum_{r=1}^m \langle v_r, b_j\rangle^3.\end{aligned} \tag{234}$$

We now aim at computing the mean and variance of $\mathcal{P}(a_1, \dots, a_P)$. Those quantities are obtained through the corresponding Hermite polynomial of P as stated in Lemma J.11. Let $\mathcal{H}(x)$ be an Hermite polynomial of degree 3. Since the Hermite basis is given by $H_0(x) = 1$, $H_{e_1}(x) = x$, $H_{e_2}(x) = x^2 - 1$ and $H_{e_3}(x) = x^3 - 3x$, for $\alpha_j, \beta_j, \gamma_j, \delta_j \in \mathbb{R}$, we have:

$$\begin{aligned}\mathcal{H}(a_1, \dots, a_P) &= \sum_{j=1}^P \alpha_j H_{e_3}(a_j) + \sum_{j=1}^P \beta_j H_{e_2}(a_j) + \gamma \sum_{j=1}^P H_{e_1}(a_j) + \delta \sum_{j=1}^P H_{e_0}(a_j) \\ &= \sum_{j=1}^P \alpha_j (a_j^3 - 3a_j) + \sum_{j=1}^P \beta_j (a_j^2 - 1) + \sum_{j=1}^P \gamma_j a_j + \sum_{j=1}^P \delta_j \\ &= \sum_{j=1}^P \alpha_j a_j^3 + \sum_{j=1}^P \beta_j a_j^2 + \sum_{j=1}^P (\gamma_j - 3\alpha_j) a_j + \sum_{j=1}^P (\delta_j - \beta_j).\end{aligned} \tag{235}$$

Since the decomposition of a polynomial in the monomial basis is unique, we can equate the coefficients of H and P and obtain:

$$\begin{cases}\alpha_j = \sigma^3 \sum_{r=1}^m \langle v_r, x\rangle^3 \\ \beta_j = 3\sigma^2 \sum_{r=1}^m \langle v_r, x\rangle^2 \langle v_r, b_j\rangle \\ \gamma_j = 3\sigma \sum_{r=1}^m \langle v_r, x\rangle \langle v_r, b_j\rangle^2 + 3\sigma^3 \sum_{r=1}^m \langle v_r, x\rangle^3 \\ \delta_j = \sum_{r=1}^m \langle v_r, b_j\rangle^3 + 3\sigma^2 \sum_{r=1}^m \langle v_r, x\rangle^2 \langle v_r, b_j\rangle\end{cases}. \tag{236}$$

By applying Lemma J.11, we get that $\mathrm{Var}[P(a)] = \sum_{j=1}^P \alpha_j^2 + \sum_{j=1}^P \beta_j^2 + \sum_{j=1}^P \gamma_j^2 \geq \sum_{j=1}^P \alpha_j^2$. By using this lower bound on the variance, the fact that $|\sum_{r=1}^m \langle v_r, x\rangle^3| \geq 1$ and Theorem J.2, we obtain

$$\mathbb{P}\left[\left|\sum_{j=1}^P \sum_{r=1}^m \langle v_r, \xi_j\rangle^3\right| \geq \epsilon\sigma^3\right] \geq 1 - O(d)\epsilon^{1/d} \tag{237}$$

Setting $\epsilon = 1/2$ in equation (237) yields the desired result.

□

J.1.3 Properties of the cube of a Gaussian

Lemma J.13. *Let $X \sim \mathcal{N}(0, \sigma^2)$. Then, X^3 is σ^3-subGaussian.*

Proof of Lemma J.13. By definition of the moment generating function, we have:

$$M_{X^3}(t) = \sum_{i=0}^\infty \frac{t^i E[X^{3i}]}{i!} = \sum_{k=0}^\infty \frac{t^{2k}\sigma^{6k}(2k-1)!!}{(2k)!} = \sum_{k=0}^\infty \frac{t^{2k}\sigma^{6k}}{2^k k!} = e^{\frac{t^2\sigma^6}{2}}.$$

□

Lemma J.14. *Let $(X_1, \dots, X_{P-1})$ be i.i.d. random variables such that $X_j \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$. Let $(w_1, \dots, w_m)$ be fixed vectors such that $w_r \in \mathbb{R}^d$. Therefore,*

$$\sum_{s=1}^{m} \sum_{j=1}^{P-1} \langle w_s, X_j \rangle^3 \text{ is } (\sigma^3 \sqrt{P-1} \sqrt{\textstyle\sum_{s=1}^m \|w_s\|_2^6}) - \text{subGaussian.}$$

Proof. We know that $\langle v_s^{(T)}, X_j \rangle \sim \mathcal{N}(0, \|v_s^{(T)}\|_2^2 \sigma^2)$. Therefore, $\langle v_s^{(T)}, X_j \rangle^3$ is the cube of a centered Gaussian. From Lemma J.13, $\langle v_s^{(T)}, X_j \rangle^3$ is $\sigma^3 \|w_s\|_2^3$-subGaussian. Using Lemma J.2, we deduce that $\sum_{j=1}^{P-1} \langle w_s, X_j \rangle^3$ is $\sqrt{P} \sigma^3 \|w_s\|_2^3$-subGaussian. Applying again Lemma J.2, we finally obtain that $\sum_{s=1}^m \sum_{j=1}^{P-1} \langle w_s, X_j \rangle^3$ is $\sigma^3 \sqrt{P-1} \sqrt{\sum_{s=1}^m \|w_s\|_2^6}$-subGaussian. □

J.2 Tensor Power Method Bound

In this subsection we establish a lemma for comparing the growth speed of two sequences of updates of the form $z^{(t+1)} = z^{(t)} + \eta C^{(t)} (z^{(t)})^2$. This technique is reminiscent of the classical analysis of the growth of eigenvalues on the (incremental) tensor power method of degree 2 and is stated in full generality in (Allen-Zhu & Li, 2020).

J.2.1 Bounds for GD

Lemma J.15. *Let $\{z^{(t)}\}_{t=0}^T$ be a positive sequence defined by the following recursions*

$$\begin{cases} z^{(t+1)} \geq z^{(t)} + m (z^{(t)})^2 \\ z^{(t+1)} \leq z^{(t)} + M (z^{(t)})^2 \end{cases},$$

where $z^{(0)} > 0$ is the initialization and $m, M > 0$.Let $\upsilon > 0$ such that $z^{(0)} \leq \upsilon$. Then, the time t_0 such that $z_t \geq \upsilon$ for all $t \geq t_0$ is:

$$t_0 \leq \frac{3}{m z^{(0)}} + \frac{8M}{m} \left\lceil \frac{\log(\upsilon / z_0)}{\log(2)} \right\rceil.$$

Proof of Lemma J.15. Let $n \in \mathbb{N}^*$. Let T_n be the time where $z^{(t)} \geq 2^n z^{(0)}$. This time exists because $z^{(t)}$ is a non-decreasing sequence. We want to find an upper bound on this time. We start with the case $n = 1$. By summing the recursion, we have:

$$z^{(T_1)} \geq z^{(0)} + m \sum_{s=0}^{T_1 - 1} (z^{(s)})^2. \tag{238}$$

We use the fact that $z^{(s)} \geq z^{(0))}$ in equation (238) and obtain:

$$T_1 \leq \frac{z^{(T_1)} - z^{(0)}}{m (z^{(0)})^2}. \tag{239}$$

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

$$z^{(T_1)} \leq z^{(T_1 - 1)} + M (z^{(T_1 - 1)})^2 \leq 2 z^{(0)} + 4M (z^{(0)})^2. \tag{240}$$

Combining equation (239) and equation (240), we get a bound on T_1.

$$T_1 \leq \frac{1}{m (z^{(0)})} + \frac{4M}{m}. \tag{241}$$

Now, let's find a bound for T_n. Starting from the recursion and using the fact that $z^{(s)} \geq 2^{n-1} z^{(0)}$ for $s \geq T_{n-1}$ we have:

$$z^{(T_n)} \geq z^{(T_{n-1})} + m \sum_{s=T_{n-1}}^{T_n - 1} (z^{(s)})^2 \geq z^{(T_{n-1})} + (2^{n-1})^2 m (z^{(0)})^2 (T_n - T_{n-1}). \tag{242}$$

On the other hand, by using $z^{(T_n-1)} \leq 2^n z^{(0)}$ we upper bound $z^{(T_n)}$ as follows.

$$z^{(T_n)} \leq z^{(T_n-1)} + M(z^{(T_n-1)})^2 \leq 2^n z^{(0)} + M 2^{2n} (z^{(0)})^2. \tag{243}$$

Besides, we know that $z^{(T_{n-1})} \geq 2^{n-1} z^{(0)}$. Therefore, we upper bound $z^{(T_n)} - z^{(T_{n-1})}$ as

$$z^{(T_n)} - z^{(T_{n-1})} \leq 2^{n-1} z^{(0)} + M 2^{2n} (z^{(0)})^2. \tag{244}$$

Combining equation (242) and equation (244) yields:

$$T_n \leq T_{n-1} + \frac{1}{2^{n-1} m (z^{(0)})} + \frac{4M}{m}. \tag{245}$$

We now sum equation (245) for $n = 2, \ldots, n$, use equation (241) and obtain:

$$T_n \leq T_1 + \frac{2}{m z^{(0)}} + \frac{4Mn}{m} \leq \frac{3}{m z^{(0)}} + \frac{4M(n+1)}{m} \leq \frac{3}{m z^{(0)}} + \frac{8Mn}{m}. \tag{246}$$

Lastly, we know that n satisfies $2^n z^{(0)} \geq \upsilon$ this implies that we can set $n = \left\lceil \frac{\log(\upsilon / z_0)}{\log(2)} \right\rceil$ in equation (246). □

Lemma J.16. *Let $\{z^{(t)}\}_{t=0}^T$ be a positive sequence defined by the following recursion*

$$\begin{cases} z^{(t)} \geq z^{(0)} + A \sum_{s=0}^{t-1} (z^{(s)})^2 - C \\ z^{(t)} \leq z^{(0)} + A \sum_{s=0}^{t-1} (z^{(s)})^2 + C \end{cases}, \tag{247}$$

where $A, C > 0$ and $z^{(0)} > 0$ is the initialization. Assume that $C = o(z^{(0)})$. Let $\upsilon > 0$ such that $z^{(0)} \leq \upsilon$. Then, the time t_0 such that $z^{(t)} \geq \upsilon$ is upper bounded as:

$$t_0 \leq 8 \left\lceil \frac{\log(\upsilon / z_0)}{\log(2)} \right\rceil + \frac{21}{(z^{(0)}) A}.$$

Proof of Lemma J.16. Let $n \in \mathbb{N}^*$. Let T_n be the time where $z^{(t)} \geq 2^{n-1} z^{(0)}$. We want to upper bound this time. We start with the case $n = 1$. We have:

$$z^{(T_1)} \geq z^{(0)} + A \sum_{s=0}^{T_1 - 1} (z^{(s)})^2 - C \tag{248}$$

By assumption, we know that $C = o(z^{(0)})$. This implies that for all $z^{(t)} \geq z^{(0)}/2$ for all $t \geq 0$. Plugging this in equation (248) yields:

$$z^{(T_1)} \geq z^{(0)} + \frac{A}{4} T_1 (z^{(0)})^2 - C \tag{249}$$

From equation (249), we deduce that:

$$T_1 \leq 4 \frac{z^{(T_1)} - z^{(0)} + C}{A (z^{(0)})^2}. \tag{250}$$

Now, we want to upper bound $z^{(T_1)} - z^{(0)}$. Using equation (247), we deduce that:

$$\begin{cases} z^{(T_1)} \geq z^{(0)} + A \sum_{s=0}^{T_1 - 1} (z^{(s)})^2 - C \\ z^{(T_1 - 1)} \leq z^{(0)} + A \sum_{s=0}^{T_1 - 2} (z^{(s)})^2 + C \end{cases}. \tag{251}$$

Combining the two equations in equation (251) yields

$$z^{(T_1)} - z^{(T_1 - 1)} \leq A (z^{(T_1 - 1)})^2 + 2C. \tag{252}$$

Since T_1 is the first time where $z^{(T_1)} \geq z^{(0)}$, we have $z^{(T_1 - 1)} \leq z^{(0)}$. Plugging this in equation (252) leads to:

$$z^{(T_1)} \leq z^{(0)} + A (z^{(0)})^2 + 2C. \tag{253}$$

Finally, using equation (253) in equation (250) and $C = o(z^{(0)})$ gives an upper bound on T_1.

$$T_1 \leq 4 + \frac{3C}{A(z^{(0)})^2} \leq 4 + \frac{3}{A(z^{(0)})}. \tag{254}$$

Now, let's find a bound for T_n. Starting from the recursion, we have:

$$\begin{cases} z^{(T_n)} \geq z^{(0)} + A\sum_{s=0}^{T_n-1}(z^{(s)})^2 - C \\ z^{(T_{n-1})} \leq z^{(0)} + A\sum_{s=0}^{T_{n-1}-1}(z^{(s)})^2 + C \end{cases}. \tag{255}$$

We substract the two equations in equation (255), use $z^{(s)} \geq 2^{n-2}$ for $s \geq T_{n-1}$ and obtain:

$$z^{(T_n)} - z^{(T_{n-1})} \geq A \sum_{s=T_{n-1}}^{T_n-1} (z^{(s)})^2 - 2C \geq 2^{2(n-2)}(z^{(0)})^2 A(T_n - T_{n-1}) - 2C. \tag{256}$$

On the other hand, from the recursion, we have the following inequalities:

$$\begin{cases} z^{(T_n)} \leq z^{(0)} + A\sum_{s=0}^{T_n-1}(z^{(s)})^2 - C \\ z^{(T_n-1)} \geq z^{(0)} + A\sum_{s=0}^{T_n-2}(z^{(s)})^2 - C \end{cases}. \tag{257}$$

We substract the two equations in equation (257), use $z^{(T_n-1)} \leq 2^{n-1}z^{(0)}$ and upper bound $z^{(T_n)}$ as follows.

$$z^{(T_n)} \leq z^{(T_n-1)} + A(z^{(T_n-1)})^2 + 2C \leq 2^{n-1}z^{(0)} + 2^{2(n-1)}A(z^{(0)})^2 + 2C. \tag{258}$$

Besides, we know that $z^{(T_{n-1})} \geq 2^{n-2}z^{(0)}$. Therefore, we upper bound $z^{(T_n)} - z^{(T_{n-1})}$ as

$$z^{(T_n)} - z^{(T_{n-1})} \leq 2^{n-2}z^{(0)} + 2^{2(n-1)}A(z^{(0)})^2 + 2C. \tag{259}$$

Combining equation (256) and equation (259) yields:

$$T_n \leq T_{n-1} + 4 + \frac{1}{2^{(n-2)}(z^{(0)})A} + \frac{4C}{2^{2(n-2)}(z^{(0)})^2 A} \tag{260}$$

We now sum equation (260) for $n = 2, \ldots, n$, use $C = o(z^{(0)})$ and then equation (254) to obtain:

$$T_n \leq T_1 + 4n + \frac{2}{(z^{(0)})A} + \frac{16C}{(z^{(0)})^2 A} \leq T_1 + 4n + \frac{18}{(z^{(0)})A} \leq 4(n+1) + \frac{21}{(z^{(0)})A}. \tag{261}$$

Lastly, we know that n satisfies $2^n z^{(0)} \geq \upsilon$ this implies that we can set $n = \left\lceil \frac{\log(\upsilon/z_0)}{\log(2)} \right\rceil$ in equation (261). □

J.2.2 Bounds for GD+M

Lemma J.17 (Tensor Power Method for momentum)**.** *Let $\gamma \in (0,1)$. Let $\{c^{(t)}\}_{t\geq 0}$ and $\{\mathcal{G}^{(t)}\}$ be positive sequences defined by the following recursions*

$$\begin{cases} \mathcal{G}^{(t+1)} = \gamma\mathcal{G}^{(t)} - \alpha^3(c^{(t)})^2, \\ c^{(t+1)} = c^{(t)} - \eta\mathcal{G}^{(t+1)} \end{cases},$$

and respectively initialized by $z^{(0)} \geq 0$ and $\mathcal{G}^{(0)} = 0$. Let $\upsilon \in \mathbb{R}$ such that $z^{(0)} \leq \upsilon$. Then, the time t_0 such that $z^{(t)} \geq \upsilon$ is:

$$t_0 = \frac{1}{1-\gamma}\left\lceil \frac{\log(\upsilon)}{\log(1+\delta)} \right\rceil + \frac{1+\delta}{\eta(1-e^{-1})\alpha^3 c^{(0)}},$$

where $\delta \in (0,1)$.

Proof of Lemma J.17. Let $\delta \in (0,1)$. We want to prove the following induction hypotheses:

1. After $T_n = \frac{n}{1-\gamma} + \sum_{j=0}^{n-2} \frac{\delta(\delta+1)^j}{\eta(1-e^{-1})\alpha^3 c^{(0)} \sum_{\tau=0}^{j} e^{-(j-\tau)}(1+\delta)^{2\tau}}$ iterations, we have:

$$-\mathcal{G}^{(T_n)} \geq (1-e^{-1})\alpha^3(c^{(0)})^2 \sum_{\tau=0}^{n-1} e^{-(n-1-\tau)}(1+\delta)^{2\tau}. \tag{TPM-1}$$

2. After $T'_n = \frac{n}{1-\gamma} + \sum_{j=0}^{n-1} \frac{\delta(\delta+1)^j}{\eta(1-e^{-1})\alpha^3 c^{(0)} \sum_{\tau=0}^{j} e^{-(j-\tau)}(1+\delta)^{2\tau}}$, we have:

$$c^{(T'_n)} \geq (1+\delta)^n c^{(0)}. \tag{TPM-2}$$

Let's first prove equation (TPM-1) and equation (TPM-2) for $n = 1$. First, by using the momentum update, we have:

$$-\mathcal{G}^{(T_1)} = (1-\gamma)\alpha^3 \sum_{\tau=0}^{T_1-1} \gamma^{T_1-1-\tau}(c^{(\tau)})^2 \geq \alpha^3(1-\gamma^{T_0})(c^{(0)})^2. \tag{262}$$

Setting $T_1 = 1/(1-\gamma)$ and using $\gamma = 1-\varepsilon$, we have $1-\gamma^{\frac{1}{1-\gamma}} = 1-\exp(\log(1-\varepsilon)/\varepsilon) = 1-e^{-1}$. Plugging this in equation (262) yields equation (TPM-1) for $n = 1$.

Regarding equation (TPM-2), we use the iterate update to have:

$$\begin{aligned} c^{(T'_1)} &= c^{(T_1)} - \eta \sum_{\tau=T_1}^{T'_1-1} \mathcal{G}^{(\tau)} \\ &\geq c^{(0)} + \eta\alpha^3(1-e^{-1})(c^{(0)})^2(T'_1 - T_1), \end{aligned} \tag{263}$$

where we used $c^{(T_1)} \geq c^{(0)}$ and equation (262) to obtain equation (263). Since $T'_1 + 1$ is the first time where $c^{(t)} \geq (1+\delta)c^{(0)}$, we further simplify equation (263) to obtain:

$$T'_1 = T_1 + \frac{\delta}{\eta\alpha^3(1-e^{-1})c^{(0)}} = \frac{1}{1-\gamma} + \frac{\delta}{\eta\alpha^3(1-e^{-1})c^{(0)}}. \tag{264}$$

We therefore obtained equation (TPM-2) for $n = 1$. Let's now assume equation (TPM-1) and equation (TPM-2) for n. We now want to prove these induction hypotheses for $n+1$. First, by using the momentum update, we have:

$$-\mathcal{G}^{(T_{n+1})} = -\gamma^{T_{n+1}-T'_n}\mathcal{G}^{(T'_n)} + (1-\gamma)\alpha^3 \sum_{\tau=T'_n}^{T_{n+1}-1} \gamma^{T_{n+1}-1-\tau}(c^{(\tau)})^2. \tag{265}$$

From equation (TPM-2) for n, we know that $c^{(t)} \geq (1+\delta)^n c^{(0)}$ for $t > T'_n$. Therefore, equation (265) becomes:

$$-\mathcal{G}^{(T_{n+1})} = -\gamma^{T_{n+1}-T'_n}\mathcal{G}^{(T'_n)} + \alpha^3(1-\gamma^{T_{n+1}-T'_n})(1+\delta)^{2n}(c^{(0)})^2. \tag{266}$$

From equation (TPM-1), we know that $-\mathcal{G}^{(T'_n)} \geq (1-e^{-1})\alpha^3(c^{(0)})^2 \sum_{\tau=0}^{n-1} e^{-(n-1-\tau)}(1+\delta)^{2\tau}$ for $t \geq T_n$. Therefore, we simplify equation (266) as:

$$\begin{aligned} -\mathcal{G}^{(T_{n+1})} \geq{}& \gamma^{T_{n+1}-T'_n}(1-e^{-1})\alpha^3(c^{(0)})^2 \sum_{\tau=0}^{n-1} e^{-(n-1-\tau)}(1+\delta)^{2\tau} \\ &+ \alpha^3(1-\gamma^{T_{n+1}-T'_n})(1+\delta)^{2n}(c^{(0)})^2. \end{aligned} \tag{267}$$

When we set T_{n+1} as in equation (TPM-1), we have $T_{n+1} - T'_n = \frac{1}{1-\gamma}$. Moreover, since $\gamma = 1-\varepsilon$, we have $\gamma^{\frac{1}{1-\gamma}} = e^{-1}$. Using these two observations, equation (267) is thus equal to:

$$\begin{aligned} -\mathcal{G}^{(T_{n+1})} &\geq (1-e^{-1})\alpha^3(c^{(0)})^2 \sum_{\tau=0}^{n-1} e^{-(n-\tau)}(1+\delta)^{2\tau} \\ &\quad + \alpha^3(1-e^{-1})(1+\delta)^{2n}(c^{(0)})^2 \\ &= (1-e^{-1})\alpha^3(c^{(0)})^2 \sum_{\tau=0}^{n} e^{-(n-\tau)}(1+\delta)^{2\tau}. \end{aligned} \tag{268}$$

We therefore proved equation (TPM-1) for $n+1$. Now, let's prove equation (TPM-2). We use the iterates update and obtain:

$$c^{(T'_{n+1})} = c^{(T_{n+1})} - \eta \sum_{\tau=T_{n+1}}^{T'_{n+1}-1} \mathcal{G}^{(\tau)}$$

$$\geq (\delta+1)^n c^{(0)} + \eta(1-e^{-1})\alpha^3 (c^{(0)})^2 \sum_{\tau=0}^{n} e^{-(n-\tau)}(1+\delta)^{2\tau}(T_{n+1} - T'_{n+1}), \tag{269}$$

where we used $c^{(T_{n+1})} \geq (\delta+1)^n c^{(0)}$ and equation (268) in the last inequality. Since $T'_{n+1}+1$ is the first time where $c^{(t)} \geq (1+\delta)^{n+1} c^{(0)}$, we further simplify equation (269) to obtain:

$$\begin{aligned} T'_{n+1} &= T_{n+1} + \frac{\delta(\delta+1)^{n-1}}{\eta(1-e^{-1})\alpha^3 (c^{(0)})^2 \sum_{\tau=0}^{n} e^{-(n-\tau)}(1+\delta)^{2\tau}} \\ &= \frac{n+1}{1-\gamma} + \sum_{j=0}^{n-1} \frac{\delta(\delta+1)^j}{\eta(1-e^{-1})\alpha^3 c^{(0)} \sum_{\tau=0}^{j} e^{-(j-\tau)}(1+\delta)^{2\tau}} \\ &+ \frac{\delta(\delta+1)^n}{\eta(1-e^{-1})\alpha^3 (c^{(0)})^2 \sum_{\tau=0}^{n} e^{-(n-\tau)}(1+\delta)^{2\tau}} \\ &= \frac{n+1}{1-\gamma} + \sum_{j=0}^{n} \frac{\delta(\delta+1)^j}{\eta(1-e^{-1})\alpha^3 c^{(0)} \sum_{\tau=0}^{j} e^{-(j-\tau)}(1+\delta)^{2\tau}}. \end{aligned} \tag{270}$$

We therefore proved equation (TPM-2) for $n+1$.

Let's now obtain an upper bound on T'_n. We have:

$$\begin{aligned} T'_n &\leq \frac{n}{1-\gamma} + \frac{\delta}{\eta(1-e^{-1})\alpha^3 c^{(0)}} \sum_{j=0}^{n-1} \frac{1}{(1+\delta)^j} \\ &\leq \frac{n}{1-\gamma} + \frac{1+\delta}{\eta(1-e^{-1})\alpha^3 c^{(0)}} := \mathscr{T}_n. \end{aligned} \tag{271}$$

Finally, we choose n such that $(1+\delta)^n \geq \upsilon$ or equivalently, $n = \left\lceil \frac{\log(\upsilon)}{\log(1+\delta)} \right\rceil$. Plugging this choice in $\mathscr{T}_n$ yields the desired bound. □

J.3 Optimization lemmas

Definition J.1 (Smooth function)**.** *Let $f\colon \mathbb{R}^{n\times d} \to \mathbb{R}$. f is β-smooth if $\|\nabla f(X) - \nabla f(Y)\|_2 \leq \beta\|X-Y\|_2$, for all $X, Y \in \mathbb{R}^{n\times d}$. A consequence of the smoothness is the inequality:*

$$f(X) \leq f(Y) + \langle \nabla f(Y), X - Y\rangle + \frac{L}{2}\|X-Y\|_2^2, \quad \textit{for all } X, Y \in \mathbb{R}^{n\times d}. \textit{Let}$$

Lemma J.18 (Descent lemma for GD)**.** *Let $f\colon \mathbb{R}^{n\times d} \to \mathbb{R}$ be a β-smooth function. Let $W^{(t+1)} \in \mathbb{R}^{n\times d}$ be an iterate of GD with learning rate $\eta \in (0, 1/L]$. Then, we have*

$$f(W^{(t+1)}) \leq f(W^{(t)}) - \frac{\eta}{2}\|\nabla f(W^{(t)})\|_2^2.$$

Proof of Lemma J.18. By applying the definition of smooth functions and the GD update, we have:

$$\begin{aligned} f(W^{(t+1)}) &\leq f(W^{(t)}) + \langle \nabla f(W^{(t)}), W^{(t+1)} - W^{(t)}\rangle + \frac{L}{2}\|W^{(t+1)} - W^{(t)}\|_2^2 \\ &= f(W^{(t)}) - \eta\|\nabla f(W^{(t)})\|_2^2 + \frac{L\eta^2}{2}\|\nabla f(W^{(t)})\|_2^2. \end{aligned} \tag{272}$$

Setting $\eta < 1/L$ in equation (272) leads to the expected result.

□

Lemma J.19 (Sublinear convergence). *Let $\mathscr{T} \geq 0$. Let $(x_t)_{t>\mathscr{T}}$ be a non-negative sequence that satisfies the recursion: $x^{(t+1)} \leq x^{(t)} - A(x^{(t)})^2$, for $A > 0$. Then, it is bounded at a time $t > \mathscr{T}$ as*

$$x^{(t)} \leq \frac{1}{A(t-\mathscr{T})}. \tag{273}$$

Proof of Lemma J.19. Let $\tau \in (\mathscr{T}, t]$. By multiplying each side of the recursion by $(x^{(\tau)}x^{(\tau+1)})^{-1}$, we get:

$$\frac{Ax^{(\tau)}}{x^{(\tau+1)}} \leq \frac{1}{x^{(\tau+1)}} - \frac{1}{x^{(\tau)}}. \tag{274}$$

Besides, the update rule indicates that $x^{(\tau)}$ is non-increasing i.e. $x^{(\tau+1)} \leq x^{(\tau)}$. Using this fact in equation (274) yields:

$$A \leq \frac{1}{x^{(\tau+1)}} - \frac{1}{x^{(\tau)}}. \tag{275}$$

Now, we sum up equation (275) for $\tau = \mathscr{T}, \dots, t-1$ and obtain:

$$A(t-\mathscr{T}) \leq \frac{1}{x^{(t)}} - \frac{1}{x^{(\mathscr{T})}} \leq \frac{1}{x^{(t)}}. \tag{276}$$

Inverting equation (276) yields the expected result. □

J.4 Other useful lemmas

Lemma J.20 (Connection between derivative and loss). *Let $a_1, \dots, a_m \in \mathbb{R}$ such that $-\delta \leq a_i \leq A$ where $A, \delta > 0$. Assume that $\sum_{i=1}^m a_i \in (C_-, C_+)$, where $C_+, C_- > 0$.Then, the following inequality holds:*

$$\frac{0.05e^{-6mA^2\delta}}{C_+\left(1+\frac{m^2\delta^2}{C_-^2}\right)} \log\left(1+e^{-\sum_{i=1}^m a_i^3}\right) \leq \frac{\sum_{i=1}^m a_i^2}{1+\exp(\sum_{i=1}^m a_i^3)} \leq \frac{20me^{6mA^2\delta}}{C_-} \log\left(1+e^{-\sum_{i=1}^m a_i^3}\right).$$

Proof of Lemma J.20. We apply Lemma J.21 to the sequence $a_i + \delta$ and obtain:

$$\begin{aligned}\frac{0.1}{C+} \log\left(1+\exp\left(-\sum_{i=1}^m (a_i+\delta)^3\right)\right) &\leq \frac{\sum_{i=1}^m (a_i+\delta)^2}{1+\exp(\sum_{i=1}^m (a_i+\delta)^3)} \\ &\leq \frac{10m}{C_-} \log\left(1+\exp\left(-\sum_{i=1}^m (a_i+\delta)^3\right)\right).\end{aligned} \tag{277}$$

We apply Lemma J.24 to further simplify equation (277).

$$\begin{aligned}&\frac{0.1e^{-\sum_{i=1}^m (3a_i^2\delta+3a_i\delta^2+\delta^3)}}{C+} \log\left(1+\exp\left(-\sum_{i=1}^m a_i^3\right)\right) \\ &\leq \frac{\sum_{i=1}^m (a_i+\delta)^2}{1+\exp(\sum_{i=1}^m (a_i+\delta)^3)} \\ &\leq \frac{10m(1+e^{-\sum_{i=1}^m (3a_i^2\delta+3a_i\delta^2+\delta^3)})}{C_-} \log\left(1+\exp\left(-\sum_{i=1}^m a_i^3\right)\right).\end{aligned} \tag{278}$$

We remark that the term inside the exponential in equation (278) can be bounded as:

$$0 \leq 2\sum_{i=1}^m a_i^2\delta \leq \sum_{i=1}^m (3a_i^2\delta - 2\delta^3) \leq \sum_{i=1}^m (3a_i^2\delta + 3a_i\delta^2 + \delta^3) \leq 6\sum_{i=1}^m a_i^2\delta \leq 6A^2m\delta. \tag{279}$$

Plugging equation (279) in equation (278) yields:

$$\begin{aligned}&\frac{0.1e^{-6mA^2\delta}}{C+}\log\left(1+\exp\left(-\sum_{i=1}^m a_i^3\right)\right)\\&\leq\frac{\sum_{i=1}^m(a_i+\delta)^2}{1+\exp(\sum_{i=1}^m(a_i+\delta)^3)}\\&\leq\frac{20m}{C_-}\log\left(1+\exp\left(-\sum_{i=1}^m a_i^3\right)\right).\end{aligned}\tag{280}$$

Lastly, we need to bound the term in the middle in equation (280). On one hand, we have:

$$\sum_{i=1}^m(a_i+\delta)^2=2\sum_{i=1}^m a_i^2+2m\delta^2\leq 2\left(1+\frac{m^2\delta^2}{\left(\sum_{i=1}^m a_i\right)^2}\right)\sum_{i=1}^m a_i^2\leq 2\left(1+\frac{m^2\delta^2}{C_-^2}\right)\sum_{i=1}^m a_i^2.\tag{281}$$

Besides, since $x\mapsto x^3$ is non-decreasing, we have the following lower bound:

$$\sum_{i=1}^m(a_i+\delta)^3\geq\sum_{i=1}^m a_i^3.\tag{282}$$

Combining equation (281) and equation (282) yields:

$$\frac{\sum_{i=1}^m(a_i+\delta)^2}{1+\exp(\sum_{i=1}^m(a_i+\delta)^3)}\leq 2\left(1+\frac{m^2\delta^2}{C_-^2}\right)\frac{\sum_{i=1}^m a_i^2}{1+\exp(\sum_{i=1}^m a_i^3)}.\tag{283}$$

On the other hand, we have:

$$\sum_{i=1}^m(a_i+\delta)^2\geq\sum_{i=1}^m a_i^2+2\delta\sum_{i=1}^m a_i\geq\sum_{i=1}^m a_i^2+2\delta C_-\geq\sum_{i=1}^m a_i^2.\tag{284}$$

Besides, using equation (279), we have:

$$\sum_{i=1}^m(a_i+\delta)^3\leq\sum_{i=1}^m a_i^3+6A^2m\delta.\tag{285}$$

Thus, using equation (284) and equation (285) yields:

$$\frac{\sum_{i=1}^m(a_i+\delta)^2}{1+\exp(\sum_{i=1}^m(a_i+\delta)^3)}\geq\frac{e^{-6mA^2\delta}\sum_{i=1}^m a_i^2}{1+\exp(\sum_{i=1}^m a_i^3)}.\tag{286}$$

Finally, we obtain the desired result by combining equation (280), equation (283) and equation (286).

□

Lemma J.21 (Connection between derivative and loss for positive sequences)**.** *Let $a_1,\ldots,a_m\in\mathbb{R}$ such that $a_i\geq 0$. Assume that $\sum_{i=1}^m a_i\in(C_-,C_+)$, where $C_+,C_->0$. Then, the following inequality holds:*

$$\frac{0.1}{C+}\log\left(1+\exp\left(-\sum_{i=1}^m a_i^3\right)\right)\leq\frac{\sum_{i=1}^m a_i^2}{1+\exp(\sum_{i=1}^m a_i^3)}\leq\frac{10m}{C_-}\log\left(1+\exp\left(-\sum_{i=1}^m a_i^3\right)\right).$$

Proof of Lemma J.21. We first remark that:

$$\begin{aligned}\frac{\sum_{i=1}^m a_i^2}{1+\exp(\sum_{i=1}^m a_i^3)}&=\frac{\left(\sum_{i=1}^m a_i^2\right)\left(\sum_{j=1}^m a_j\right)}{\left(1+\exp(\sum_{i=1}^m a_i^3)\right)\left(\sum_{j=1}^m a_j\right)}\\&=\frac{\sum_{i=1}^m a_i^3+\sum_{i=1}^m\sum_{j\neq i}a_i^2a_j}{\left(1+\exp(\sum_{i=1}^m a_i^3)\right)\left(\sum_{j=1}^m a_j\right)}.\end{aligned}\tag{287}$$

Upper bound. We upper bound equation (287) by successively applying $\sum_{i=1}^n a_i > C_-$ and $a_i > 0$ for all i:

$$\begin{aligned}\frac{\sum_{i=1}^m a_i^2}{1+\exp(\sum_{i=1}^m a_i^3)} &\leq \frac{\sum_{i=1}^m a_i^3 + \sum_{i=1}^m \sum_{j\neq i} a_i^2 a_j}{C_- \left(1+\exp(\sum_{i=1}^m a_i^3)\right)} \\ &\leq \frac{\sum_{i=1}^m a_i^3 + \sum_{i=1}^m \sum_{j=1}^m a_i^2 a_j}{C_- \left(1+\exp(\sum_{i=1}^m a_i^3)\right)} \end{aligned} \tag{288}$$

where we used $a_i > 0$ for all i in equation (287). By applying the rearrangement inequality to equation (288), we obtain:

$$\frac{\sum_{i=1}^m a_i^2}{1+\exp(\sum_{i=1}^m a_i^3)} \leq \frac{m}{C_-} \frac{\sum_{i=1}^m a_i^3}{1+\exp(\sum_{i=1}^m a_i^3)}. \tag{289}$$

We obtain the final bound by applying Lemma J.22 to equation (289).

Lower bound. We lower bound equation (287) by using $\sum_{i=1}^n a_i \leq C_+$ and $\sum_{i=1}^m \sum_{j\neq i} a_i^2 a_j$:

$$\begin{aligned}\frac{\sum_{i=1}^m a_i^2}{1+\exp(\sum_{i=1}^m a_i^3)} &\geq \frac{\sum_{i=1}^m a_i^3 + \sum_{i=1}^m \sum_{j\neq i} a_i^2 a_j}{C_+ \left(1+\exp(\sum_{i=1}^m a_i^3)\right)} \\ &\geq \frac{\sum_{i=1}^m a_i^3}{C_+ \left(1+\exp(\sum_{i=1}^m a_i^3)\right)}. \end{aligned} \tag{290}$$

We obtain the final bound by applying Lemma J.22 to equation (290). □

Lemma J.22 (Connection between derivative and loss)**.** *Let $x > 0$. Then, we have:*

$$0.1 \log(1+\exp(-x)) \leq \mathfrak{S}(x) \leq 10 \log(1+\exp(-x)) \tag{291}$$

Lemma J.23. *Let $(x^{(t)})_{t\geq 0}$ be a non-negative sequence. Let $A > 0$. Assume that $\sum_{\tau=0}^T x^{(\tau)} \leq A$. Then, there exists a time $\mathscr{T} \in [T]$ such that $x^{(\mathscr{T})} \leq A/T$.*

Proof of Lemma J.23. Assume by contradiction that for all $\tau \in [T]$, $x^{(\tau)} > A/T$. By summing up x^τ, we obtain $\sum_{\tau=0}^T x^{(\tau)} > A$. This contradicts the assumption that $\sum_{\tau=0}^T x^{(\tau)} \leq A$.

□

Lemma J.24 (Log inequalities)**.** *Let $x, y > 0$. Then, the following inequalities holds:*

1. *Assume that $y \leq x$. We have:*

$$\log(1+xy) \leq (1+y)\log(1+x).$$

2. *Assume $y < 1$. We have:*

$$y \log(1+x) \leq \log(1+xy).$$

Proof of Lemma J.24. We first remark that:

$$\begin{aligned}\log(1+xy) - \log(1+x) &= \log\left(\frac{1+xy}{1+x}\right) \\ &= \log\left(1 + \frac{x(y-1)}{1+x}\right). \end{aligned} \tag{292}$$

From equation (292), we deduce an upper bound as:

$$\log(1+xy) - \log(1+x) \leq \log\left(1 + \frac{x(y+1)}{1+x}\right). \tag{293}$$

Successively using the inequalities $\log(1+x) \leq x$ and $\frac{x}{1+x} \leq \log(1+x)$ for $x > -1$ in equation (293) yields:

$$\log(1+xy) - \log(1+x) \leq (1+y)\frac{x}{1+x} \leq (1+y)\log(1+x).$$

This proves item 1 of the Lemma. Let's now prove item 2. Using $a^z \leq 1 + (a-1)z$ for $z \in (0,1)$ and $a \geq 1$, we know that:

$$(1+x)^y \leq 1 + xy. \tag{294}$$

Since $\log$ is non-decreasing, applying $\log$ to equation (294) proves item 2.

We now prove item 3. □