DISSECTING ADAPTIVE METHODS IN GANS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive methods are a crucial component widely used for training generative adversarial networks (GANs). While there has been some work to pinpoint the "marginal value of adaptive methods" in standard tasks, it remains unclear why they are still critical for GAN training. In this paper, we formally study how adaptive methods help train GANs; inspired by the grafting method proposed in Agarwal et al. (2020), we separate the magnitude and direction components of the Adam updates, and graft them to the direction and magnitude of SGDA updates respectively. By considering an update rule with the magnitude of the Adam update and the normalized direction of SGD, we empirically show that the adaptive magnitude of Adam is key for GAN training. This motivates us to have a closer look at the class of normalized stochastic gradient descent ascent (nSGDA) methods in the context of GAN training. We propose a synthetic theoretical framework to compare the performance of nSGDA and SGDA for GAN training with neural networks. We prove that in that setting, GANs trained with nSGDA recover all the modes of the true distribution, whereas the same networks trained with SGDA (and any learning rate configuration) suffer from mode collapse. The critical insight in our analysis is that normalizing the gradients forces the discriminator and generator to be updated at the same pace. We also experimentally show that for several datasets, Adam's performance can be recovered with nSGDA methods.

1 INTRODUCTION

Adaptive algorithms have become a key component in training modern neural network architectures in various deep learning tasks. Minimization problems that arise in natural language processing (Vaswani et al., 2017), fMRI (Zbontar et al., 2018), or min-max problems such as generative adversarial networks (GANs) (Goodfellow et al., 2014) almost exclusively use adaptive methods, and it has been empirically observed that Adam (Kingma & Ba, 2014) yields a solution with better generalization than stochastic gradient descent (SGD) in such problems (Choi et al., 2019). Several works have attempted to explain this phenomenon in the minimization setting. Common explanations are that adaptive methods train faster (Zhou et al., 2018), escape flat "saddle-point"–like plateaus faster (Orvieto et al., 2021), or handle heavy-tailed stochastic gradients better (Zhang et al., 2019; Gorbunov et al., 2022). However, much less is known about why adaptive methods are so critical for solving min-max problems such as GANs.

Several previous works attribute this performance to the superior convergence speed of adaptive methods. For instance, Liu et al. (2019) show that an adaptive variant of Optimistic Gradient Descent (Daskalakis et al., 2017) converges faster than stochastic gradient descent ascent (SGDA) for a class of non-convex, non-concave min-max problems. However, contrary to the minimization setting, convergence to a stationary point is not required to obtain satisfactory GAN performance. Mescheder et al. (2018) empirically shows that popular architectures such as Wasserstein GANs (WGANs) (Arjovsky et al., 2017) do not always converge, and yet they produce realistic images. We support this observation with our own experiments in Section 2. Our findings motivate the central question in this paper: *what factors of Adam contribute to better quality solutions over SGDA when training GANs?*

In this paper, we investigate why GANs trained with adaptive methods outperform those trained using SGDA. Directly analyzing Adam is challenging due to the highly non-linear nature of its gradient oracle and its path-dependent update rule. Inspired by the grafting approach in (Agarwal et al., 2020), we disentangle the adaptive magnitude and direction of Adam and show evidence that an algorithm using the adaptive magnitude of Adam and the direction of SGDA (which we call Ada-nSGDA) recovers the performance of Adam in GANs. Our contributions are as follows:

- In Section 2, we present the Ada-nSGDA algorithm.We emprically show that the adaptive magnitude in Ada-nSGDA stays within a constant range and does not heavily fluctuate which motivates the focus on normalized SGDA (nSGDA) which only contains the direction of SGDA.
- In Section 3, we prove that for a synthetic dataset consisting of two modes, a model trained with SGDA suffers from *mode collapse* (producing only a single type of output), while a model trained with nSGDA does not. This provides an explanation for why GANs trained with nSGDA outperform those trained with SGDA.
- In Section 4, we empirically confirm that nSGDA mostly recovers the performance of Ada-nSGDA when using different GAN architectures on a wide range of datasets.

Our key theoretical insight is that when using SGDA and any step-size configuration, either the generator G or discriminator D updates much faster than the other. By normalizing the gradients as done in nSGDA, D and G are forced to update at the same speed throughout training. The consequence is that whenever D learns a mode of the distribution, G learns it right after, which makes both of them learn all the modes of the distribution ~~separately~~ at the same pace.

1.1 Related work

Adaptive methods in games optimization. Several works designed adaptive algorithms and analyzed their convergence to show their benefits relative to SGDA e.g. in variational inequality problems, Gasnikov et al. (2019); Antonakopoulos et al. (2019); Bach & Levy (2019); Antonakopoulos et al. (2020); Liu et al. (2019); Barazandeh et al. (2021). Heusel et al. (2017) show that Adam locally converges to a Nash equilibrium in the regime where the step-size of the discriminator is much larger than the one of the generator. Our work differs as we do not focus on the convergence properties of Adam, but rather on the fit of the trained model to the *true* (and not empirical) data distribution.

Statistical results in GANs. Early works studied whether GANs memorize the training data or actually learn the distribution (Arora et al., 2017; Liang, 2017; Feizi et al., 2017; Zhang et al., 2017; Arora et al., 2018; Bai et al., 2018; Dumoulin et al., 2016). Some works explained GAN performance through the lens of optimization. Lei et al. (2020); Balaji et al. (2021) show that GANs trained with SGDA converge to a global saddle point when the generator is one-layer neural network and the discriminator is a specific quadratic/linear function. Our contribution differs as i) we construct a setting where SGDA converges to a locally optimal min-max equilibrium but still suffers from mode collapse, and ii) we have a more challenging setting since we need at least a degree-3 discriminator to learn the distribution, which is discussed in Section 3.

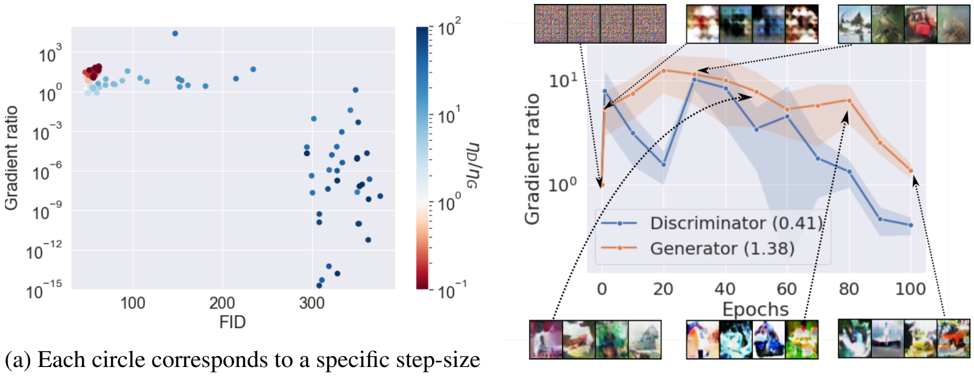
Normalized gradient descent. Introduced by Nesterov (1984), normalized gradient descent has been widely used in minimization problems. Normalizing the gradient remedies the issue of iterates being stuck in flat regions such as spurious local minima or saddle points (Hazan et al., 2015; Levy, 2016). Normalized gradient descent methods outperforms their non-normalized counterparts in multi-agent coordination (Cortés, 2006) and deep learning tasks (Cutkosky & Mehta, 2020). Our work considers the min-max setting and shows that nSGDA outperforms SGDA as it forces discriminator and generator to update at same rate.

1.2 Background

Generative adversarial networks. Given a training set sampled from some target distribution $\mathcal{D}$, a GAN learns to generate new data from this distribution. The architecture is comprised of two networks: a generator that maps points in the latent space $\mathcal{D}_z$ to samples of the desired distribution, and a discriminator which evaluates these samples by comparing them to samples from $\mathcal{D}$. More formally, the generator is a mapping $G_{\mathcal{V}} \colon \mathbb{R}^k \to \mathbb{R}^d$ and the discriminator is a mapping $D_{\mathcal{W}} \colon \mathbb{R}^d \to \mathbb{R}$, where $\mathcal{V}$ and $\mathcal{W}$ are their corresponding parameter sets. To find the optimal parameters of these two networks, one must solve a min-max optimization problem of the form

$$\min_{\mathcal{V}} \max_{\mathcal{W}} \mathbb{E}_{X\sim p_{data}}[\log(D_{\mathcal{W}}(X))] + \mathbb{E}_{z\sim p_z}[\log(1 - D_{\mathcal{W}}(G_{\mathcal{V}}(z)))] := f(\mathcal{V}, \mathcal{W}), \tag{GAN}$$

where p_{data} is the distribution of the training set, p_z the latent distribution, $G_{\mathcal{V}}$ the generator and $D_{\mathcal{W}}$ the discriminator. Contrary to minimization problems where convergence to a local minimum is *required* for high generalization, we empirically verify that most of the well-performing GANs do not converge to a stationary point.



(a) Each circle corresponds to a specific step-size configuration η_D/η_G. The best-performing models have step-size ratios between 10^{-1} and 1, and do not converge. As η_D/η_G increases, the models perform worse but get closer to an equilibrium.

(b) shows that during training, the gradient ratio of a well-performing GAN approximately stays constant to 1. We also display the images produced by the model during training.

Figure 1: Gradient ratio against FID score (a) and number of epochs (b) obtained with DCGAN on CIFAR-10. This ratio is equal to $\|\text{grad}_G^{(t)}\|_2/\|\text{grad}_G^{(0)}\|_2 + \|\text{grad}_D^{(t)}\|_2/\|\text{grad}_D^{(0)}\|_2$, where $\text{grad}_G^{(t)}$ (resp. $\text{grad}_D^{(t)}$) and $\text{grad}_G^{(0)}$ (resp. $\text{grad}_D^{(0)}$) are the current and initial gradients of G (resp. D). Note that $\|\cdot\|_2$ refers to the sum of all the parameters norm in a network. For all the plots, the models are trained for 100 epochs using a batch-size 64. For (b), the results are averaged over 5 seeds.

Convergence and performance are decorrelated in GANs. We support this observation through the following experiment. We train a DCGAN (Radford et al., 2015) using Adam and set up the step-sizes for G and D as η_D, η_G, respectively. Note that D is usually trained faster than G i.e. $\eta_D \geq \eta_G$. Figure 1a displays the GAN convergence measured by the ratio of gradient norms and the GAN's performance measured in FID score (Heusel et al., 2017). We observe that when η_D/η_G is close to 1, the algorithm does not converge, but the model produces high-quality samples. On the other hand, when $\eta_D/\eta_G \gg 1$, the model converges to an equilibrium; a similar statement has been proved by Jin et al. (2020) and Fiez & Ratliff (2020) in the case of SGDA. However, the trained GAN produces low-quality solutions at this equilibrium, so simply comparing the convergence speed of adaptive methods and SGDA cannot explain the performance obtained with adaptive methods.

SGDA and adaptive methods. The most simple algorithm to solve the min-max (GAN) is SGDA, which is defined as follows:

$$\mathcal{W}^{(t+1)} = \mathcal{W}^{(t)} + \eta_D \mathbf{M}_{\mathcal{W},1}^{(t)}, \quad \mathcal{V}^{(t+1)} = \mathcal{V}^{(t)} - \eta_G \mathbf{M}_{\mathcal{V},1}^{(t)}, \tag{1}$$

where $\mathbf{M}_{\mathcal{W},1}^{(t)}, \mathbf{M}_{\mathcal{V},1}^{(t)}$ are the first-order momentum gradients as defined in Algorithm 1. While this method has been used in the first GANs (Radford et al., 2015), most modern GANs are trained with adaptive methods such as Adam (Kingma & Ba, 2014). The definition of this algorithm for game optimizations is given in Algorithm 1. The hyperparameters $\beta_1, \beta_2 \in [0, 1)$ control the weighting of the exponential moving average of the first and second-order moments. In many deep-learning tasks, practitioners have found that setting $\beta_2 = 0.9$ works for most problem settings. Additionally, it has been empirically observed that having no momentum (i.e., $\beta_1 \approx 0$) is optimal for many popular GAN architectures (Karras et al., 2020; Brock et al., 2018). Therefore, we only consider the case where $\beta_1 = 0$.

Optimizers such as Adam (Algorithm 1) are *adaptive* since they use a step-size for each parameter that is different than the magnitude of the gradient $\mathbf{g}_{\mathcal{Y}}^{(t)}$ for that parameter up to some constant (such as the global learning rate), and this step-size updates while training the model. There are three components that makes the adaptive update differ from the standard SGDA update: 1) the *adaptive normalization* by $\|\mathbf{g}_{\mathcal{Y}}^{(t)}\|_2$, 2) the *change of direction* from $\mathbf{g}_{\mathcal{Y}}^{(t)}/\|\mathbf{g}_{\mathcal{Y}}^{(t)}\|_2$ to $\mathbf{A}_{\mathcal{Y}}^{(t)}/\|\mathbf{A}_{\mathcal{Y}}^{(t)}\|_2$ and 3) *adaptive scaling* by $\|\mathbf{A}_{\mathcal{Y}}^{(t)}\|_2$. In summary, the steps from the standard to the adaptive update, are:

$$\mathbf{g}_{\mathcal{Y}}^{(t)} \xrightarrow[\times 1/\|\mathbf{g}_{\mathcal{Y}}^{(t)}\|_2]{\text{normalization}} \mathbf{g}_{\mathcal{Y}}^{(t)}/\|\mathbf{g}_{\mathcal{Y}}^{(t)}\|_2 \xrightarrow{\text{change of direction}} \mathbf{A}_{\mathcal{Y}}^{(t)}/\|\mathbf{A}_{\mathcal{Y}}^{(t)}\|_2 \xrightarrow[\times \|\mathbf{A}_{\mathcal{Y}}^{(t)}\|_2]{\text{adaptive scaling}} \mathbf{A}_{\mathcal{Y}}^{(t)} \tag{2}$$

The three components are entangled and it remains unclear how they contribute to the superior performance of adaptive methods relative to SGDA in GANs.

Algorithm 1 Adam (Kingma & Ba, 2014) for games. All operations on vectors are element-wise.

```
**Input**: initial points $\mathcal{W}^{(0)}, \mathcal{V}^{(0)}$, step-size schedules $\{(\eta_G^{(t)}, \eta_D^{(t)})\}$ , hyperparameters $\{\beta_1, \beta_2, \varepsilon\}$.
Initialize $\mathbf{M}^{(0)}_{\mathcal{W},1}, \mathbf{M}^{(0)}_{\mathcal{W},2}, \mathbf{M}^{(0)}_{\mathcal{V},1}$ and $\mathbf{M}^{(0)}_{\mathcal{V},2}$ to zero.
**for** $t = 0 \ldots T-1$ **do**
    Receive stochastic gradients $\mathbf{g}_{\mathcal{W}}^{(t)}, \mathbf{g}_{\mathcal{V}}^{(t)}$ evaluated at $\mathcal{W}^{(t)}$ and $\mathcal{V}^{(t)}$.
    Update for $\mathcal{Y} \in \{\mathcal{W}, \mathcal{V}\}$: $\mathbf{M}^{(t+1)}_{\mathcal{Y},1} = \beta_1 \mathbf{M}^{(t)}_{\mathcal{Y},1} + \mathbf{g}_{\mathcal{Y}}^{(t)}$ and $\mathbf{M}^{(t+1)}_{\mathcal{Y},2} = \beta_2 \mathbf{M}^{(t)}_{\mathcal{Y},2} + {\mathbf{g}_{\mathcal{Y}}^{(t)}}^2$.
    Compute gradient oracles for $Y \in \{V, W\}$: $\mathbf{A}_{\mathcal{Y}}^{(t+1)} = \mathbf{M}^{(t+1)}_{\mathcal{Y},1} / \sqrt{\mathbf{M}^{(t+1)}_{\mathcal{Y},2} + \varepsilon}$.
    Update: $\mathcal{W}^{(t+1)} = \mathcal{W}^{(t)} + \eta_D^{(t)} \mathbf{A}_{\mathcal{W}}^{(t+1)}, \qquad \mathcal{V}^{(t+1)} = \mathcal{V}^{(t)} - \eta_G^{(t)} \mathbf{A}_{\mathcal{V}}^{(t+1)}$.
```

2 nSGDA as a model to analyze Adam in GANs

In this section, we show that normalized stochastic gradient descent-ascent (nSGDA) is a suitable proxy to study the learning dynamics of Adam.

To decouple the normalization, change of direction, and adaptive scaling in Adam, we adopt the grafting approach proposed by Agarwal et al. (2020). At each iteration, we compute stochastic gradients, pass them to two optimizers $\mathcal{A}_1, \mathcal{A}_2$ and make a grafted step that combines the *magnitude* of $\mathcal{A}_1$'s step and *direction* of $\mathcal{A}_2$'s step. We focus on the optimizer defined by grafting the Adam magnitude onto the SGDA direction, which corresponds to omitting the *change of direction* in (2):

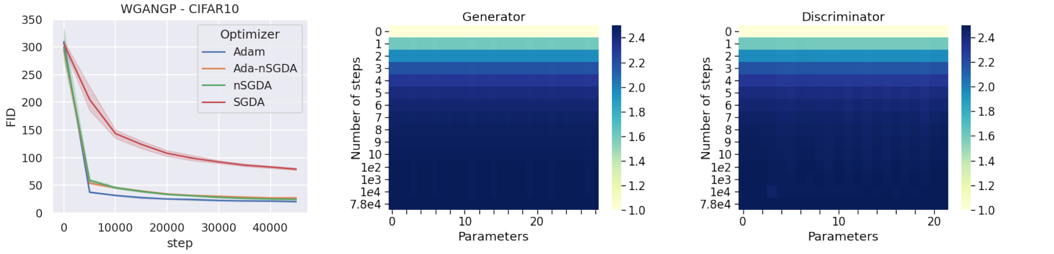
$$\mathcal{W}^{(t+1)} = \mathcal{W}^{(t)} + \eta_D \|\mathbf{A}_{\mathcal{W}}^{(t)}\|_2 \frac{\mathbf{g}_{\mathcal{W}}^{(t)}}{\|\mathbf{g}_{\mathcal{W}}^{(t)}\|_2 + \varepsilon}, \quad \mathcal{V}^{(t+1)} = \mathcal{V}^{(t)} - \eta_G \|\mathbf{A}_{\mathcal{V}}^{(t)}\|_2 \frac{\mathbf{g}_{\mathcal{V}}^{(t)}}{\|\mathbf{g}_{\mathcal{V}}^{(t)}\|_2 + \varepsilon}, \tag{3}$$

where $\mathbf{A}_{\mathcal{V}}^{(t)}, \mathbf{A}_{\mathcal{W}}^{(t)}$ are the Adam gradient oracles as in Algorithm 1 and $\boldsymbol{g}_{\mathcal{V}}^{(t)}, \boldsymbol{g}_{\mathcal{W}}^{(t)}$ the stochastic gradients. We refer to this algorithm as *Ada-nSGDA* (combining the Adam magnitude and SGDA direction). There are two natural implementations for nSDGA. In the *layer-wise* version, $\mathcal{Y}^{(t)}$is a single parameter group (typically a layer in a neural network), and the updates are applied to each group. In the *global* version, $\mathcal{Y}^{(t)}$ contains all of the model's weights.

In Fig. 2a, we see that Ada-nSGDA and Adam appear to have similar learning dynamics in terms of the FID score. Both Adam and Ada-nSGDA significantly outperform SGDA as well as AdaDir, which is the alternate case of (3) where we instead graft the magnitude of the SGDA update to the direction of the Adam update. AdaDir diverged after a single step so we omit it in Fig. 2. These results show that the *adaptive scaling* and *normalization* components are sufficient to recover the performance of Adam, suggesting that Ada-nSGDA is a valid proxy for Adam

A natural question that arises is how the total adaptive magnitude varies during training. We empirically investigate this by tracking the layer-wise adaptive magnitudes of the Adam gradient oracles when training a GAN with Ada-nSGDA, and summarize our key findings here (see Section 4 for complete experimental details). We first train a WGAN-GP (Arjovsky et al., 2017) model, and find that the adaptive magnitude is bound within a constant range, and that all the layers have approximately the same adaptive magnitude (Fig. 2 (b,c)). This suggests that the *adaptive scaling* component is constant (in expectation) and motivates the use of *nSGDA*, corresponding to Ada-nSGDA with a constant *adaptive scaling* factor. We then train a WGAN-GP model with nSGDA and we find that nSGDA mostly recovers the FID score obtained by Ada-nSGDA (Fig. 2a).

We also validate this observation for more complicated GAN architectures by repeating this study on StyleGAN2 (Karras et al., 2019). We find that the adaptive magnitudes also vary within a constant range, but each layer has its own constant scaling factor. Thus, training StyleGAN2 with nSGDA and a global normalization fails, but training with nSGDA with a different constant step-size for each layer yields a performance that mostly recovers that of Ada-nSGDA (Fig 5). These results suggest that the schedule of the *adaptive scaling* is not central in the success of Ada-nSGDA in GANs. Instead, adaptive methods are successful because they *normalize* the gradients for each layer, which allows for more balanced updates between G and D as we will show in Section 3. We conduct more experiments in Section 4 and in Appendix A.



(a)

(b)

(c)

Figure 2: (a) shows the FID training curve for a WGAN-GP ResNet, averaged over 5 seeds. We see that Ada-nSGDA and nSGDA have very similar performance to Adam for a WGAN-GP. (b, c) displays the fluctuations of Ada-nSGDA adaptive magnitude. We plot the ratio $\|\mathbf{A}_{\mathcal{Y}}^{(t)}\|_2/\|\mathbf{A}_{\mathcal{Y}}^{(0)}\|_2$ for each of the generator's (b) and discriminator's (c) layers. At early stages, this ratio barely increases and remains constant after 10 steps.

3 Why does nSGDA perform better than SGDA in GANs?

In Section 2, we empirically showed that the most important component of the adaptive magnitude is the *normalization*, and that nSGDA (an algorithm consisting of this component alone) is sufficient to recover most of the performance of Ada-nSGDA (and by extension, Adam). Our goal is to construct a dataset and model where we can prove that a model trained with nSGDA generates samples from the true training distribution while SGDA fails. To this end, we consider a dataset where the underlying distribution consists of two modes, defined as vectors u_1, u_2, that are slightly correlated (see Assumption 1) and consider the standard GANs' training objective. We show that a GAN trained with SGDA using any reasonable[1] step-size configuration suffers from *mode collapse* (Theorem 3.1); it only outputs samples from a single mode which is a weighted average of u_1 and u_2. Conversely, nSGDA-trained GANs learn the two modes separately (Theorem 3.2).

Notation We set the GAN 1-sample loss $L_{\mathcal{V},\mathcal{W}}^{(t)}(X,z) = \log(D_{\mathcal{W}}^{(t)}(X)) + \log(1 - D_{\mathcal{W}}^{(t)}(G_{\mathcal{V}}^{(t)}(z)))$. We denote $\mathbf{g}_{\mathcal{Y}}^{(t)} = \nabla_{\mathcal{Y}} L_{\mathcal{V},\mathcal{W}}^{(t)}(X,z)$ as the 1-sample stochastic gradient. We use the asymptotic complexity notations when defining the different constants e.g. $\mathrm{poly}(d)$ refers to any polynomial in the dimension d, $\mathrm{polylog}(d)$ to any polynomial in $\log(d)$, and $o(1)$ to a constant $\ll d$. We denote $a \propto b$ for vectors a and b in $\mathbb{R}^d$ if there is a positive scaling factor $c > 0$ such that $\|a - cb\|_2 = o(\|b\|_2)$.

3.1 Setting

In this section, we present the setting to sketch our main results in Theorem 3.1 and Theorem 3.2. We first define the distributions for the training set and latent samples, and specify our GAN model and the algorithms we analyze to solve (GAN). Note that for many assumptions and theorems below, we present informal statements which are sufficient to capture the main insights. The precise statements can be found in Appendix B.

Our synthetic theoretical framework considers a bimodal data distribution with two correlated modes:

Assumption 1 (p_{data} structure)**.** *Let $\gamma = \frac{1}{\mathrm{polylog}(d)}$. We assume that the modes are correlated. This means that $\langle u_1, u_2 \rangle = \gamma > 0$ and the generated data point X is either $X = u_1$ or $X = u_2$.*

Next, we define the latent distribution p_z that $G_{\mathcal{V}}$ samples from and maps to p_{data}. Each sample from p_z consists of a data-point z that is a binary-valued vector $z \in \{0,1\}^{m_G}$, where m_G is the number of neurons in $G_{\mathcal{V}}$, and has non-zero support, i.e. $\|z\|_0 \geq 1$. Although the typical choice of a latent distributions in GANs is either Gaussian or uniform, we choose p_z to be a binary distribution because it models the weights' distribution of a hidden layer of a deep generator; Allen-Zhu & Li (2021) argue that the distributions of these hidden layers are sparse, non-negative, and non-positively correlated. We now make the following assumptions on the coefficients of z:

Assumption 2 (p_z structure)**.** *Let $z \sim p_z$. We assume that with probability $1 - o(1)$, there is only one non-zero entry in z. The probability that the entry $i \in [m_G]$ is non-zero is $\Pr[z_i = 1] = \Theta(1/m_G)$.*

In Assumption 2, the output of $G_{\mathcal{V}}$ is only made of one mode with probability $1 - o(1)$. This avoids summing two of the generator's neurons, which may cause mode collapse.

[1] Reasonable simply means that the learning rates are bounded to prevent the training from diverging.

To learn the target distribution p_{data}, we use a linear generator $G_{\mathcal{V}}$ with m_G neurons and a non-linear neural network with m_D neurons:

$$G_{\mathcal{V}}(z) = Vz = \sum_{i=1}^{m_G} v_i z_i \,, \qquad D_{\mathcal{W}}(X) = \text{sigmoid}\Big(a \sum_{i=1}^{m_D} \langle w_i, X\rangle^3 + \frac{b}{\sqrt{d}}\Big). \tag{4}$$

where $V = [v_1^\top, v_2^\top, \cdots, v_{m_G}^\top] \in \mathbb{R}^{m_G \times d}$, $z \in \{0,1\}^{m_G}$, $W = [w_1^\top, \ldots, w_{m_D}^\top] \in \mathbb{R}^{m_D \times d}$, and $a, b \in \mathbb{R}$. Intuitively, $G_{\mathcal{V}}$ outputs linear combinations of the modes v_i. We choose a cubic activation as it is the smallest monomial degree for the discriminator's non-linearity that is sufficient for the generator to recover the modes u_1, u_2.[2]

We now state the SGDA and nSGDA algorithms used to solve the GAN training problem (GAN). For simplicity, we set the batch-size to 1. The resultant update rules for SGDA and nSGDA are:[3]

SGDA: at each step $t > 0$, sample $X \sim p_{data}$ and $z \sim p_z$ and update as

$$\mathcal{W}^{(t+1)} = \mathcal{W}^{(t)} + \eta_D \mathbf{g}_{\mathcal{W}}^{(t)}, \quad \mathcal{V}^{(t+1)} = \mathcal{V}^{(t)} - \eta_G \mathbf{g}_{\mathcal{V}}^{(t)}, \tag{5}$$

nSGDA: at each step $t > 0$, sample $X \sim p_{data}$ and $z \sim p_z$ and update as

$$\mathcal{W}^{(t+1)} = \mathcal{W}^{(t)} + \eta_D \frac{\mathbf{g}_{\mathcal{W}}^{(t)}}{\|\mathbf{g}_{\mathcal{W}}^{(t)}\|_2}, \quad \mathcal{V}^{(t+1)} = \mathcal{V}^{(t)} - \eta_G \frac{\mathbf{g}_{\mathcal{V}}^{(t)}}{\|\mathbf{g}_{\mathcal{V}}^{(t)}\|_2}. \tag{6}$$

Compared to the versions of SGDA and Ada-nSGDA that we introduced in Section 2, we have the same algorithms except that we set $\beta_1 = 0$ and omit ε in (5) and (6). Note that since there is only one layer in the neural networks we study in this paper, the global-wise and layer-wise versions of nSGDA are actually the same. Lastly, we detail how to set the optimization parameters for SGDA and nSGDA in (5) and (6).

Parametrization 3.1 (Informal). *When running SGDA and nSGDA on (GAN), we set:*

*– **Initialization**: $b^{(0)} = 0$, and $a^{(0)}$, $w_i^{(0)} (i \in [m_D])$, $v_j^{(0)} (j \in [m_G])$ are initialized with a Gaussian with small variance.*

*– **Number of iterations**: we run SGDA for $t \leq T_0$ iterations where T_0 is the first iteration such that the algorithm converges to an approximate first order local minimum. For nSGDA, we run for $T_1 = \tilde{\Theta}(1/\eta_D)$ iterations.*

*– **Step-sizes**: For SGDA, $\eta_D, \eta_G \in (0, \frac{1}{\text{poly}(d)})$ can be arbitrary. For nSGDA, $\eta_D \in (0, \frac{1}{\text{poly}(d)}]$, and η_G is slightly smaller than η_D.*

*– **Over-parametrization**: For SGDA, $m_D, m_G = \text{polylog}(d)$ are arbitrarily chosen i.e. m_D may be larger than m_G or the opposite. For nSGDA, we set $m_D = \log(d)$ and $m_G = 2\log(d)$.*

Our theorem holds when running SGDA for any (polynomially) possible number of iterations; after T_0 steps, the gradient becomes inverse polynomially small and SGDA essentially stops updating the parameters. Additionally, our setting allows any step-size configuration for SGDA i.e. larger, smaller, or equal step-size for D compared to G. Note that our choice of step-sizes for nSGDA is the one used in practice, i.e. η_D slightly larger than η_G.

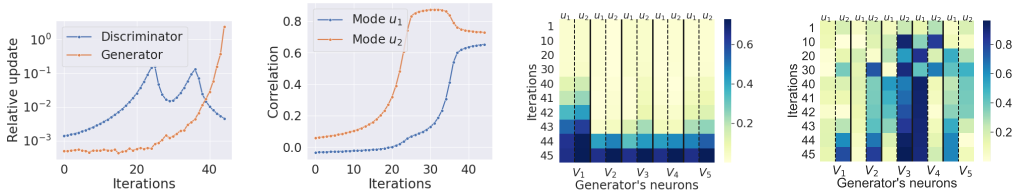
3.2 Main results

We state our main results on the performance of models trained using SGDA (5) and nSGDA (6). We show that nSGDA learns the modes of the distribution p_{data} while SGDA does not.

Theorem 3.1 (Informal). *Consider a training dataset and a latent distribution as described above and let Assumption 1 and Assumption 2 hold. Let T_0, η_G, η_D and the initialization be as defined in*

[2] Li & Dou (2020) show that when using linear or quadratic activations, the generator can fool the discriminator by only matching the first and second moments of p_{data}.

[3] In the nSGDA algorithm defined in (3), the step-sizes were time-dependent. Here, we assume for simplicity that the step-sizes $\eta_D, \eta_G > 0$ are *constant*.



(a) Relative gradients (b) D weight correlation (c) $\mathcal{V}$ learned via SGDA (d) $\mathcal{V}$ learned via nSGDA

Figure 3: (a) shows the relative gradient updates for SGDA. D first updates its weights while G does not move until iteration 20, then G moves its weights. (b) shows the correlation for one neuron of D (with maximal correlation to u_2 at initialization) with the modes u_1, u_2 during the learning process of SGDA. (c, d) shows the correlations of the neurons of G with the modes when trained with SGDA and nSGDA respectively. This shows that for SGDA (c), the model ultimately learns the weighted average $u_1 + u_2$. For nSGDA, we see from (d) that one of the neurons (V_4) is highly correlated with u_1 and another one (V_3) is correlated with u_2.

Parametrization 3.1. Let t be such that $t \leq T_0$. Run SGDA on (GAN) for t iterations with step-sizes η_G, η_D. Then, with probability at least $1 - o(1)$, the generator outputs for all $z \in \{0,1\}^{m_G}$:

$$G_{\mathcal{V}}^{(t)}(z) \propto \begin{cases} u_1 + u_2 & \textit{if } \eta_D \geq \eta_G \\ \xi^{(t)}(z) & \textit{otherwise} \end{cases}, \tag{7}$$

where $\xi^{(t)}(z) \in \mathbb{R}^d$ is some vector that is not correlated to any of the modes. Formally, $\forall \ell \in [2]$, $\cos(\xi^{(t)}(z), u_\ell) = o(1)$ for all $z \in \{0,1\}^{m_G}$.

A formal proof can be found in Appendix G. Theorem 3.1 indicates that when training with SGDA and any step-size configuration, the generator either does not learn the modes at all ($G_{\mathcal{V}}^{(t)}(z) = \xi^{(t)}(z)$) or learns an average of the modes ($G_{\mathcal{V}}^{(t)}(z) \propto u_1 + u_2$). The theorem holds *for any* time $t \leq T_0$ which is the iteration where SGDA converges to an approximate first-order locally optimal min-max equilibrium. Conversely, nSGDA succeeds in learning the two modes separately:

Theorem 3.2 (Informal)**.** *Consider a training dataset and a latent distribution as described above and let Assumption 1 and Assumption 2 hold. Let T_1, η_G, η_D and the initialization as defined in Parametrization 3.1. Run nSGDA on (GAN) for T_1 iterations with step-sizes η_G, η_D. Then, the generator learns both modes u_1, u_2 i.e., for $\ell \in \{1, 2\}$,*

$$\Pr_{z \sim p_z}[G_{\mathcal{V}}^{(T_1)}(z) \propto u_\ell] \quad \textit{is non-negligible}. \tag{8}$$

A formal proof can be found in Appendix I. Theorem 3.2 indicates that when we train a GAN with nSGDA in the regime where the discriminator updates slightly faster than the generator (as done in practice), the generator successfully learns the distribution containing the direction of both modes.

We implement the setting introduced in Subsection 3.1 and validate Theorem 3.1 and Theorem 3.2 in Fig. 3. Fig. 3a displays the relative update speed $\eta \|\mathbf{g}_{\mathcal{Y}}^{(t)}\|_2 / \|\mathcal{Y}^{(t)}\|_2$, where $\mathcal{Y}$ corresponds to the parameters of either D or G. Fig. 3b shows the correlation $\langle w_i^{(t)}, u_\ell \rangle / \|w_i^{(t)}\|_2$ between *one* of D's neurons and a mode u_ℓ and Fig. 3c the correlation $\langle v_j^{(t)}, u_\ell \rangle / \|v_j^{(t)}\|_2$ between G's neurons and u_ℓ. We discuss the interpretation of these plots to the next section.

Why does SGDA suffer from mode collapse and nSGDA learn the modes?

We now explain why SGDA suffers from mode collapse, which corresponds to the case where $\eta_D \geq \eta_G$. Our explanation relies on the interpretation of Figs. 3a, 3b, and 3c, and on the updates around initialization that are defined as followed. There exists $i \in [m_D]$ such that D's update is

$$\mathbb{E}[w_i^{(t+1)} | w_i^{(t)}] \approx w_i^{(t)} + \eta_D \sum_{l=1}^{2} \mathbb{E}[\langle w_i^{(t)}, u_l \rangle^2] u_l \,. \tag{9}$$

Thus, the weights of D receive gradients directed by u_1 and u_2. On the other hand, the weights of G at early stages receive gradients directed by $w_j^{(t)}$:

$$v_i^{(t+1)} \approx v_i^{(t)} + \eta_G \sum_j \langle v_i^{(t)}, w_j^{(t)} \rangle^2 w_j^{(t)}. \tag{10}$$

We observe that the learning process in Figs. 3a & 3b has three distinct phases. In the first phase (iterations 1-20), D learns one of the modes (u_1 or u_2) of p_{data} (Fig. 3b) and G barely updates its weights (Fig. 3a). In the second phase (iterations 20-40), D learns the weighted average $u_1 + u_2$ (Fig. 3b) while G starts moving its weights (Fig. 3a). In the final phase (iterations 40+), G learns $u_1 + u_2$ (Fig. 3c) from D. In more detail, the learning process is described as follows:

Phase 1 : At initialization, $w_j^{(0)}$ and $v_j^{(0)}$ are small. Assume w.l.o.g. that $\langle w_i^{(0)}, u_2 \rangle > \langle w_i^{(0)}, u_1 \rangle$. Because of the $\langle w_i^{(t)}, u_l \rangle^2$ in front of u_2 in (9), the parameter $w_i^{(t)}$ gradually grows its correlation with u_2 (Fig. 3b) and D's gradient norm thus increases (Fig. 3a). While $\|w_j^{(t)}\| \ll 1 \, \forall j$, we have that $v_i^{(t)} \approx v_i^{(0)}$ (Fig. 3a).

Phase 2: D has learned u_2. Because of the sigmoid in the gradient of $w_i^{(t)}$ (that was negligible during Phase 1) and $\langle u_1, u_2 \rangle = \gamma > 0$, $w_i^{(t)}$ now mainly receives updates with direction u_2. Since G did not update its weights yet, the min-max problem (GAN) is approximately just a minimization problem with respect to D's parameters. Since the optimum of such a problem is the weighted average $u_1 + u_2$, $w_j^{(t)}$ slowly converges to this optimum. Meanwhile, $v_i^{(t)}$ start to receive some significant signal (Fig. 3a) but mainly learn the direction $u_1 + u_2$ (Fig. 3c), because $w_j^{(t)}$ is aligning with this direction.

Phase 3: The parameters of G only receive gradient directed by $u_1 + u_2$. The norm of its relative updates stay large and D only changes its last layer terms (slope a and bias b).

In contrast to SGDA, nSGDA ensures that G and D always learn at the same speed with the updates:

$$w_i^{(t+1)} \approx w_i^{(t)} + \eta_D \frac{\langle w_i^{(t)}, X \rangle^2 X}{\|\langle w_i^{(t)}, X \rangle^2 X\|_2}, \text{ and } v_i^{(t+1)} \approx v_i^{(t)} + \eta_G \frac{\sum_j \langle w_j^{(t)}, v_i^{(t)} \rangle^2 w_j^{(t)}}{\|\sum_j \langle w_j^{(t)}, v_i^{(t)} \rangle^2 w_j^{(t)}\|_2} \qquad (11)$$

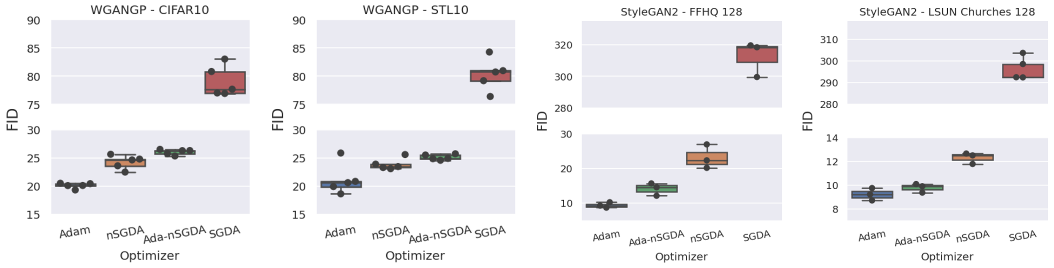
No matter how large $\langle w_i^{(t)}, X \rangle$ is, G still learns at the same speed with D. There is a tight window (iteration 25, Fig. 3b) where only one neuron of D is aligned with u_1. This is when G can also learn to generate u_1 by "catching up" to D at that point, which avoids mode collapse.

4 Numerical performance of nSGDA

In Section 2, we presented the Ada-nSGDA algorithm (3) which corresponds to "grafting" the Adam magnitude onto the SGDA direction. In Section 3, we construct a dataset and GAN model where we prove that a GAN trained with nSGDA can generate examples from the true training distribution, while a GAN trained with SGDA fails due to mode collapse. We now provide more experiments comparing nSGDA and Ada-nSGDA with Adam on real GANs and datasets.

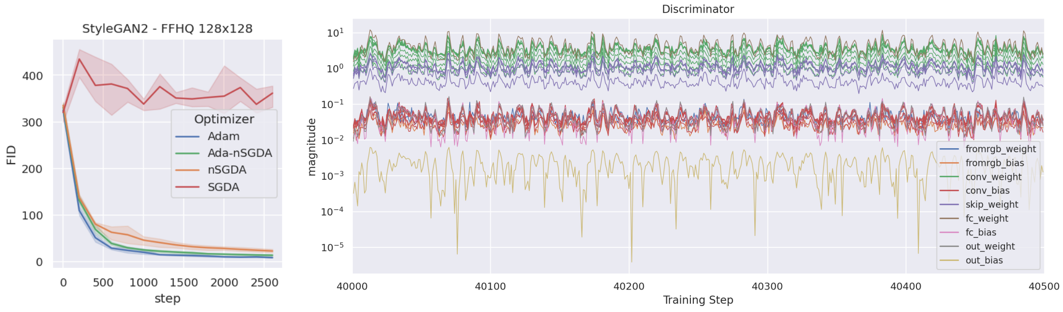
We train a ResNet WGAN with gradient penalty on CIFAR-10 (Krizhevsky et al., 2009) and STL-10 (Coates et al., 2011) with Adam, Ada-nSDGA, SGDA, and nSGDA with a fixed learning rate as done in Section 3. We use the default architectures and training parameters specified in Gulrajani et al. (2017) ($\lambda_{GP} = 10$, $n_{dis} = 5$, learning rate decayed linearly to 0 over 100k steps). We also train a StyleGAN2 model (Karras et al., 2020) on FFHQ (Karras et al., 2019) and LSUN Churches (Yu et al., 2016) (both resized to 128×128 pixels) with Adam, Ada-nSGDA, SGDA, and nSGDA. We use the recommended StyleGAN2 hyperparameter configuration for this resolution (batch size = 32, $\gamma = 0.1024$, map depth = 2, channel multiplier = 16384). We use the Fréchet Inception distance (FID) (Heusel et al., 2017) to quantitatively assess the performance of the model. For each optimizer, we conduct a coarse log-space sweep over step sizes and optimize for FID. We train the WGAN-GP models for 2880 thousand images (kimgs) on CIFAR-10 and STL-10 (45k steps with a batch size of 64), and the StyleGAN2 models for 2600 kimgs on FFHQ and LSUN Churches. We average our results over 5 seeds for the WGAN-GP ResNets, and over 3 seeds for the StyleGAN2 models due to the computational cost associated with training GANs.

WGAN-GP Figures 4a and 4b validates the conclusions on WGAN-GP from Section 2. We find that both Ada-nSGDA and nSGDA mostly recover the performance of Adam, with nSGDA obtaining a final FID of ∼2-3 points lower than Ada-nSGDA. As discussed in Section 2, such performance is possible because the adaptive magnitude stays within a constant range. In contrast, models trained with SGDA consistently perform significantly worse, with final FID scores $4\times$ larger than Adam.



(a) (b) (c) (d)

Figure 4: (a, b) are the final FID scores (5 seeds) for a ResNet WGAN-GP model trained for 45k steps on CIFAR-10 and STL-10 respectively. (c, d) are the final FID scores (3 seeds) for a StyleGAN2 model trained for 2600kimgs on FFHQ and LSUN Churches respectively. We use the same constant layer scaling in (d) for nSGDA as that in (c), which was found by tracking the layer-wise adaptive step-sizes.



(a) (b)

Figure 5: (a) shows the FID curve for the StyleGAN2 model trained for 2600 kimgs on FFHQ. (b) shows the fluctuations of the Ada-nSGDA adaptive magnitude for each layer over an arbitrary slice of 500 training steps for the Discriminator. The layers are grouped by common types, e.g. Conv weights and biases, etc.). We find that although the magnitude for each layer fluctuates, the fluctuations are bounded to some fixed range for each layer. We show similar behaviour for the Generator in the Appendix.

StyleGAN2 Figures 4c and 4d show the final FID scores when training a StyleGAN2. We find that Ada-nSGDA recovers most of the performance of Adam, but one difference with WGAN-GP is that nSGDA does not work if we use the same global learning rate for each layer. As discussed in Section 2, nSGDA with a different (but constant) step-size for each layer *does* work, and is able to mostly recover Ada-nSGDA's performance (Fig. 5a). To choose the scaling for each layer, we train StyleGAN2 with Ada-nSGDA on FFHQ-128, track the layer-wise adaptive magnitudes, and take the mean of these magnitudes over the training run (for each layer). Figure 5b shows that the fluctuations for each layer are bound to a constant range, validating our assumption of constant step-sizes. Additionally, the same scaling obtained from training FFHQ seems to work for different datasets; we used it to train StyleGAN2 with nSGDA on LSUN Churches-128 and recovered similar performance to training on this dataset with Ada-nSGDA (Fig. 4d).

5 Conclusion

Our work addresses the question of which mechanisms in adaptive methods are critical for training GANs, and why they outperform non-adaptive methods. We empirically show that Ada-nSGDA, an algorithm composed of the adaptive magnitude of Adam and the direction of SGD, recovers most of the performance of Adam. We further decompose the adaptive magnitude into two components: normalization, and adaptive step-size. We then show that the adaptive step size is roughly constant (bounded fluctuations) for multiple architectures and datasets. This empirically indicates that the normalization component of the adaptive magnitude is the key mechanism of Ada-nSGDA, and motivates the study of nSGDA; we verify that it too recovers the performance of Ada-nSGDA. Having shown that nSGDA is a good proxy for a key mechanism for adaptive methods, we then construct a setting where we proved that nSGDA –thanks to its balanced updates– recovers the modes of the true distribution while SGDA fails to do it. The key insight from our theoretical analysis is that the ratio of the update of D and G must be close to 1 during training in order to recover the modes of the distribution. This matches the experimental setting with nSGDA, as we find that global norm of the parameter updates for both D and G are almost equal for optimal choices of learning rates.

References

Naman Agarwal, Rohan Anil, Elad Hazan, Tomer Koren, and Cyril Zhang. Disentangling adaptive gradient methods from learning rates. *arXiv preprint arXiv:2002.11803*, 2020.

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *CoRR*, abs/2012.09816, 2020. URL https://arxiv.org/abs/2012.09816.

Zeyuan Allen-Zhu and Yuanzhi Li. Forward super-resolution: How can gans learn hierarchical generative models for real-world distributions. *arXiv preprint arXiv:2106.02619*, 2021.

Kimon Antonakopoulos, Veronica Belmega, and Panayotis Mertikopoulos. An adaptive mirror-prox method for variational inequalities with singular operators. *Advances in Neural Information Processing Systems*, 32:8455–8465, 2019.

Kimon Antonakopoulos, E Veronica Belmega, and Panayotis Mertikopoulos. Adaptive extra-gradient methods for min-max optimization and games. *arXiv preprint arXiv:2010.12100*, 2020.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning*, pp. 224–232. PMLR, 2017.

Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do gans learn the distribution? some theory and empirics. In *International Conference on Learning Representations*, 2018.

Francis Bach and Kfir Y Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *Conference on Learning Theory*, pp. 164–194. PMLR, 2019.

Yu Bai, Tengyu Ma, and Andrej Risteski. Approximability of discriminators implies diversity in gans. *arXiv preprint arXiv:1806.10586*, 2018.

Yogesh Balaji, Mohammadmahdi Sajedi, Neha Mukund Kalibhat, Mucong Ding, Dominik Stöger, Mahdi Soltanolkotabi, and Soheil Feizi. Understanding overparameterization in generative adversarial networks. *arXiv preprint arXiv:2104.05605*, 2021.

Babak Barazandeh, Davoud Ataee Tarzanagh, and George Michailidis. Solving a class of non-convex min-max games using adaptive momentum methods. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3625–3629. IEEE, 2021.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Dami Choi, Christopher J Shallue, Zachary Nado, Jaehoon Lee, Chris J Maddison, and George E Dahl. On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446*, 2019.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Jorge Cortés. Finite-time convergent gradient flows with applications to network consensus. *Automatica*, 42(11):1993–2000, 2006.

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

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

Soheil Feizi, Farzan Farnia, Tony Ginart, and David Tse. Understanding gans: the lqg setting. *arXiv preprint arXiv:1710.10793*, 2017.

Tanner Fiez and Lillian Ratliff. Gradient descent-ascent provably converges to strict local minmax equilibria with a finite timescale separation. *arXiv preprint arXiv:2009.14820*, 2020.

AV Gasnikov, PE Dvurechensky, FS Stonyakin, and AA Titov. An adaptive proximal method for variational inequalities. *Computational Mathematics and Mathematical Physics*, 59(5):836–841, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Eduard Gorbunov, Marina Danilova, David Dobre, Pavel Dvurechensky, Alexander Gasnikov, and Gauthier Gidel. Clipped stochastic methods for variational inequalities with heavy-tailed noise. *arXiv preprint arXiv:2206.01095*, 2022.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

Elad Hazan, Kfir Y Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. *arXiv preprint arXiv:1507.02030*, 2015.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pp. 4880–4889. PMLR, 2020.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

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.

Qi Lei, Jason Lee, Alex Dimakis, and Constantinos Daskalakis. Sgd learns one-layer networks in wgans. In *International Conference on Machine Learning*, pp. 5799–5808. PMLR, 2020.

Kfir Y Levy. The power of normalization: Faster evasion of saddle points. *arXiv preprint arXiv:1611.04831*, 2016.

Yuanzhi Li and Zehao Dou. Making method of moments great again?–how can gans learn distributions. *arXiv preprint arXiv:2003.04033*, 2020.

Tengyuan Liang. How well can generative adversarial networks learn densities: A nonparametric view. *arXiv preprint arXiv:1712.08244*, 2017.

Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. *arXiv preprint arXiv:1912.11940*, 2019.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pp. 3481–3490. PMLR, 2018.

Y.E. Nesterov. Minimization methods for nonsmooth convex and quasiconvex functions. *Econ. Mat. Met.*, 20:519–531, 01 1984.

Antonio Orvieto, Jonas Kohler, Dario Pavllo, Thomas Hofmann, and Aurelien Lucchi. Vanishing curvature and the power of adaptive methods in randomly initialized deep networks, 2021.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop, 2016.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *arXiv preprint arXiv:1912.03194*, 2019.

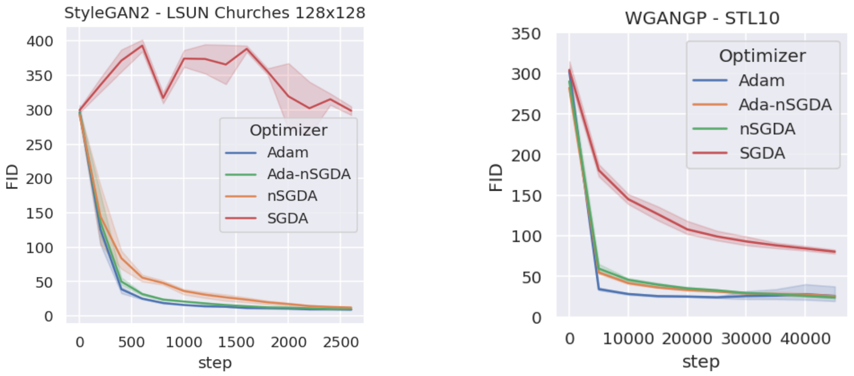
Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in gans. *arXiv preprint arXiv:1711.02771*, 2017.

Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

A Additional experiments

A.1 Experiments with StyleGAN2 and WGAN-GP

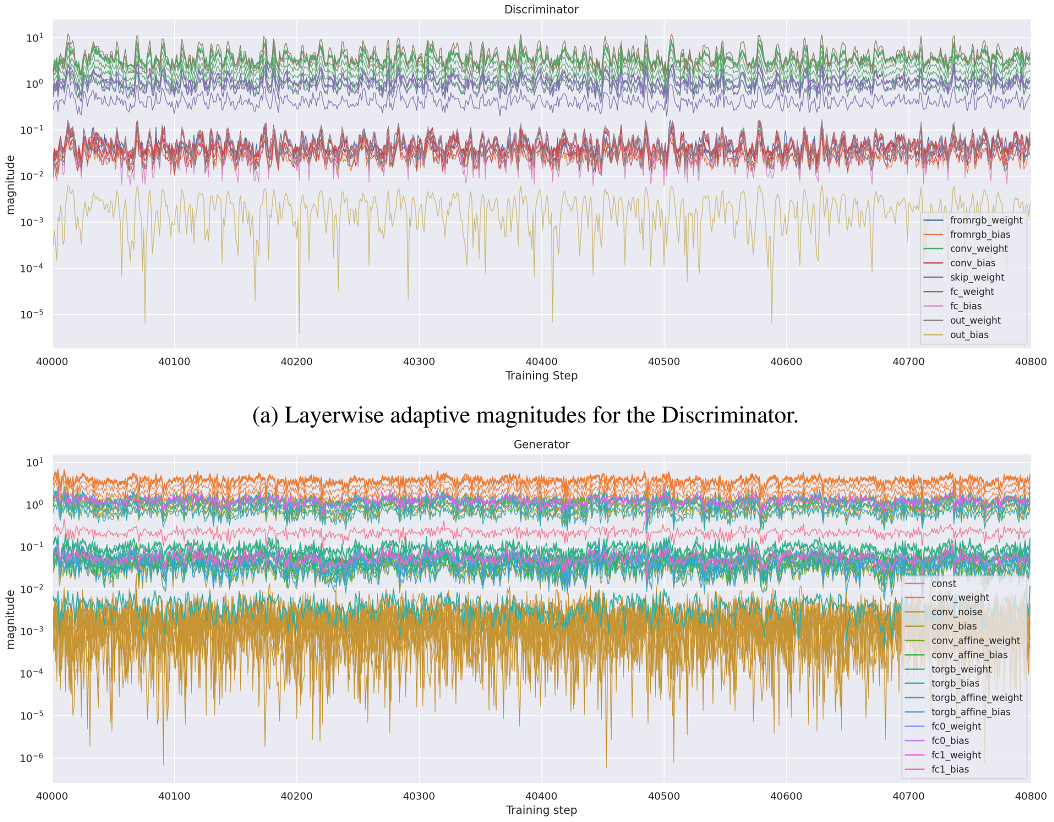
In this section, we put additional curves and images produced by WGAN and StyleGAN2.



(a)

(b)

Figure 6: (a) is the FID curve of StyleGAN2 on LSUN-Churches and (b) the FID curve of WGAN on STL-10. These complement the figures of Section 4.



(a) Layerwise adaptive magnitudes for the Discriminator.

(b) Layerwise adaptive magnitudes for the Generator.

Figure 7: The fluctuations of the Ada-nSGDA adaptive magnitudes for each layer over an arbitrary slice of 800 training steps for the Disciminator (a) and Generator (b). The layers are grouped by common types, e.g. Conv weights and biases, etc.). We find that although the magnitude for each layer fluctuates, the fluctuations are bounded to some fixed range for each layer.

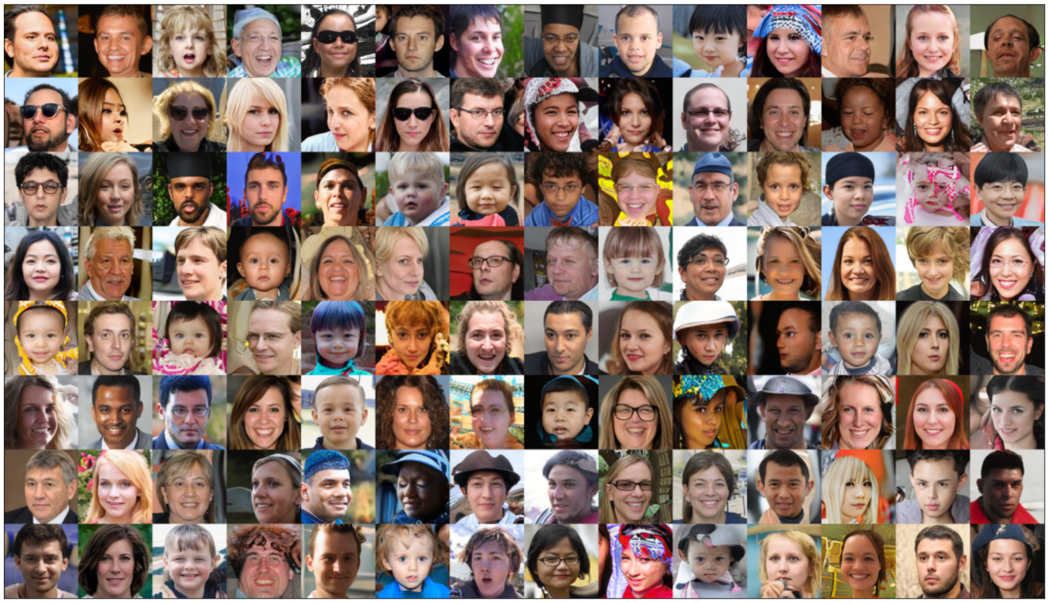


Figure 8: Images generated by a StyleGAN2 model trained with Adam for 2600 kimgs on FFHQ 128. Note that this is not convergence.

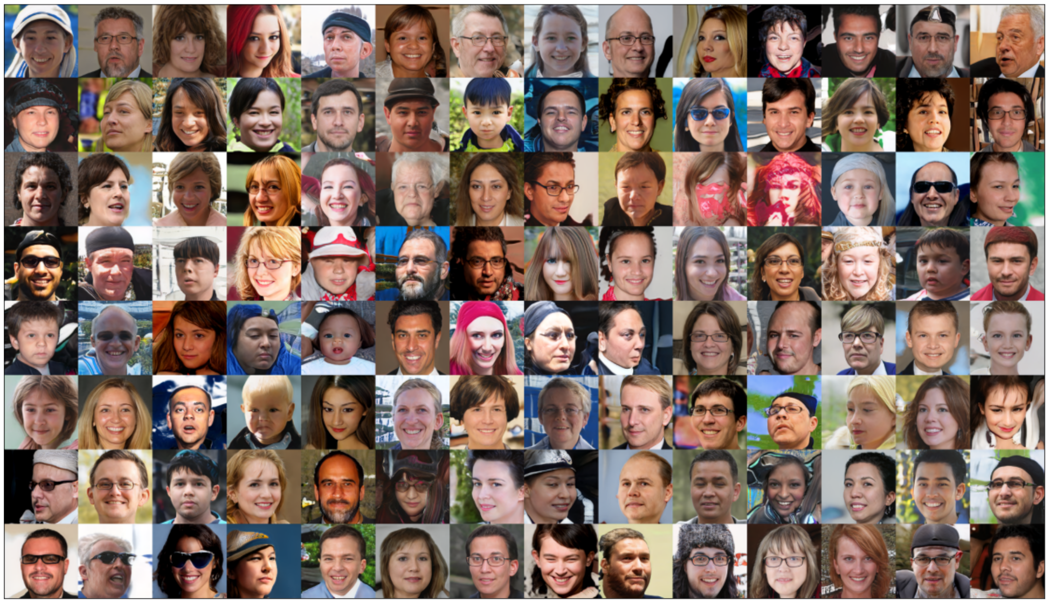


Figure 9: Images generated by a StyleGAN2 model trained with Ada-nSDGDA for 2600 kimgs on FFHQ 128. Note that this is not convergence.

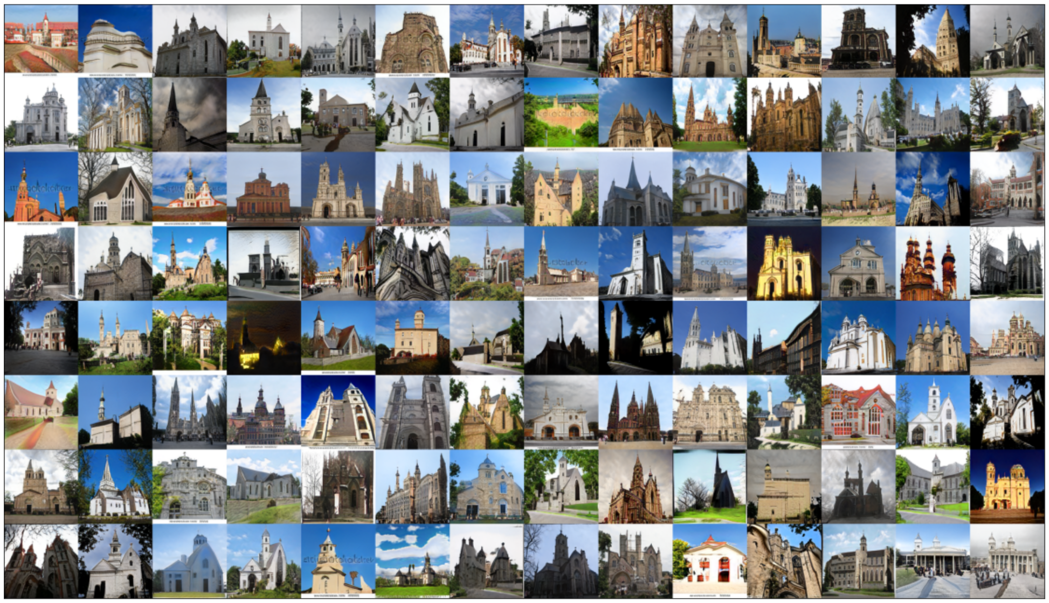


Figure 10: Images generated by a StyleGAN2 model trained with Adam for 2600 kimgs on LSUN Churches 128. Note that this is not convergence.

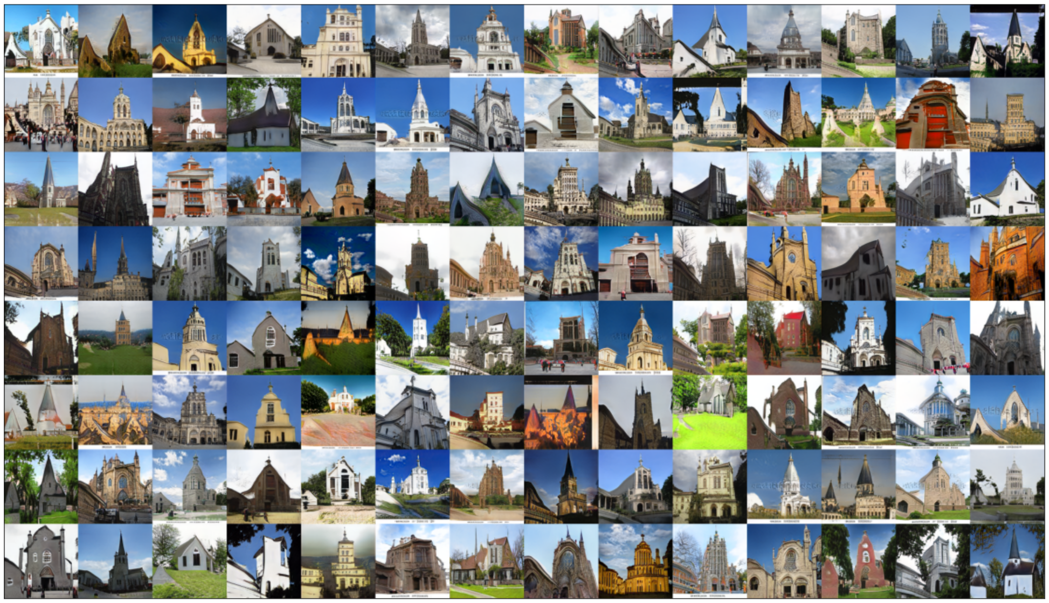
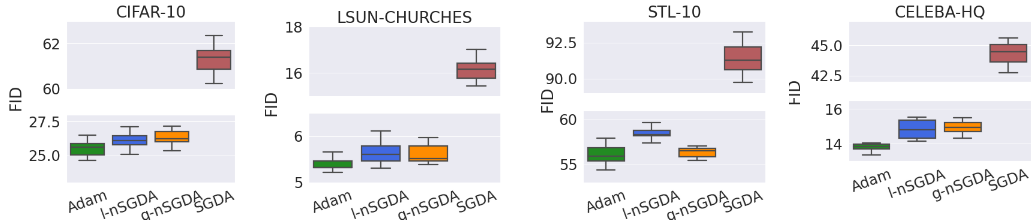


Figure 11: Images generated by a StyleGAN2 model trained with Ada-nSDGDA for 2600 kimgs on LSUN Churches 128. Note that this is not convergence.

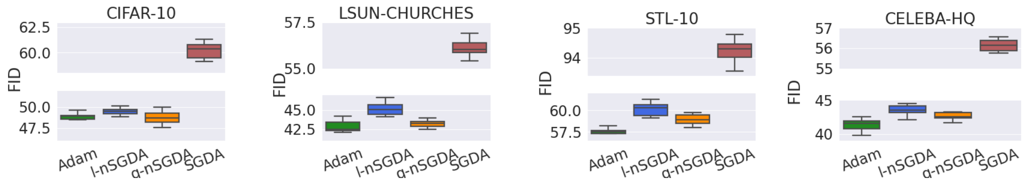
A.2 Experiments with DCGAN

This section shows that experimental results obtained in Section 4 are also valid for other architectures such as DCGAN. Indeed, we observe that nSGDA methods compete with Adam and nSGDA work when the batch size is small. In this section, lnSGDA refers to the layer-wise nSGDA and gnSGDA to the global nSGDA.



(a) CIFAR-10 (b) LSUN Churches (c) STL-10 (d) Celeba-HQ

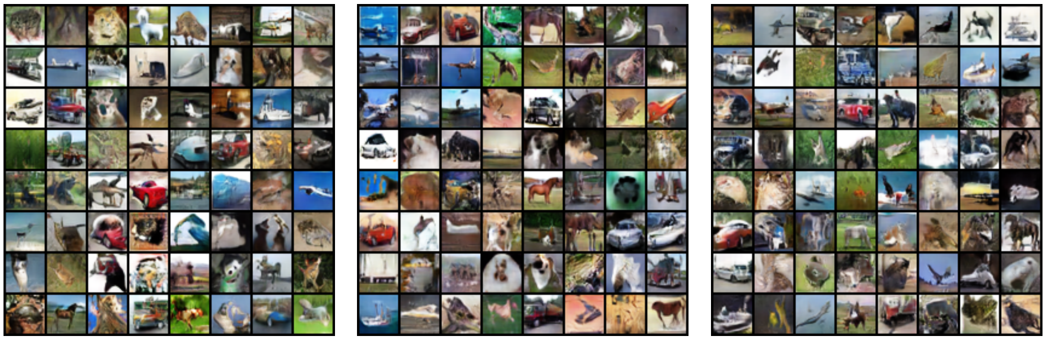
Figure 12: FID scores obtained when training a Resnet WGAN-GP using Adam, l-nSGDA, g-nSGDA, and SGD on different datasets. In all these datasets, l-nSGDA, g-nSGDA and Adam perform approximately as well. SGDA performs much worse.



(a) CIFAR-10 (b) LSUN Churches (c) STL-10 (d) Celeba-HQ

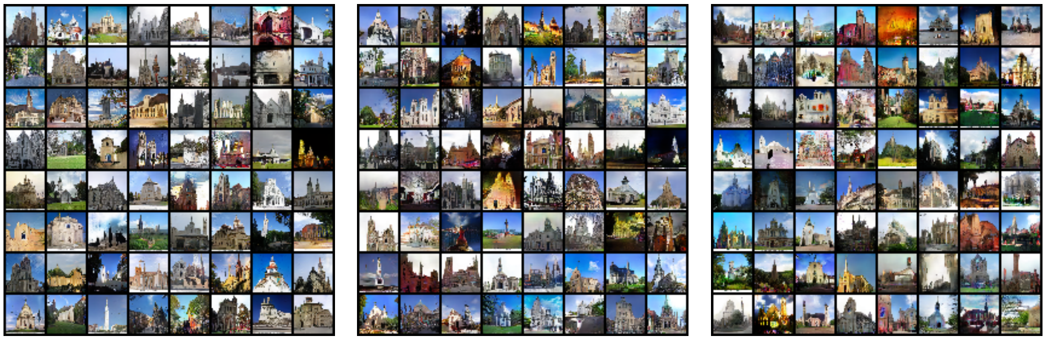
Figure 13: FID scores obtained when training a DCGAN using Adam, lnSGDA, gnSGDA and SGD on different datasets. In all these datasets, lnSGDA, gnSGDA and Adam perform approximately as well. As expected, SGDA performs much worse than the other optimizers. The models are trained with batch-size 64 –which is the usual batch-size used for DCGAN.

In this section, we display the images obtained when training the Resnet WGAN-GP model from Section 4.



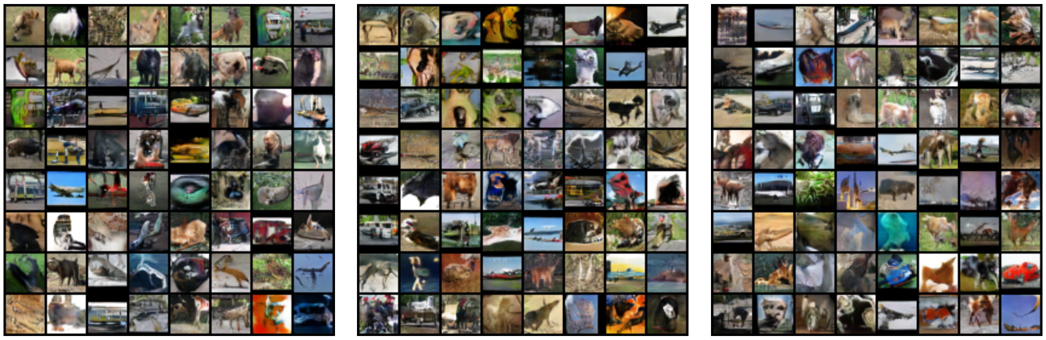
(a) Adam (b) lnSGDA (c) gnSGDA

Figure 14: CIFAR-10 images generated by a Resnet WGAN-GP model



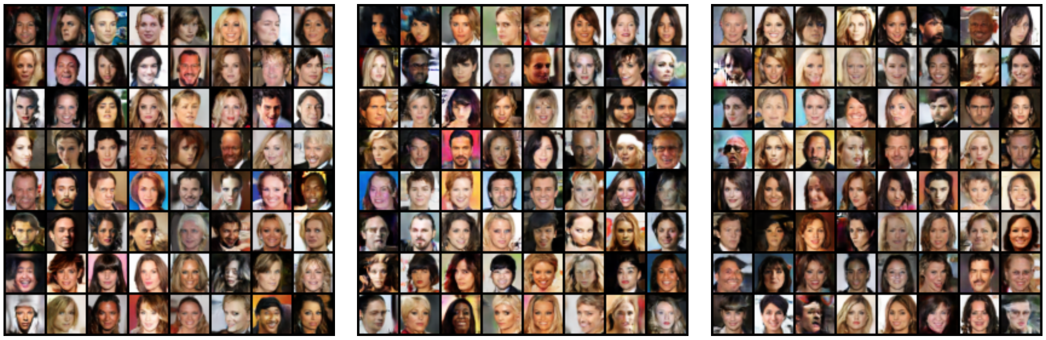
(a) Adam (b) lnSGDA (c) gnSGDA

Figure 15: LSUN-Churches images generated by a Resnet WGAN-GP model



(a) Adam (b) lnSGDA (c) gnSGDA

Figure 16: STL-10 images generated by a Resnet WGAN-GP model



(a) Adam (b) lnSGDA (c) gnSGDA

Figure 17: Celeba-HQ images generated by a Resnet WGAN-GP model

B Technical statements in the theory section

In this section, we provide the technical version of the statements made in Section 3.

B.1 Setting

The distribution p_{data} we consider is more general than Assumption 1 in the main paper.

Assumption 3 (Data structure)**.** *Let $\gamma = \frac{1}{\mathrm{polylog}(d)}$. The coefficients s_1, s_2 and modes u_1, u_2 of the distribution p_{data} respect one of the following conditions:*

1. *Correlated modes: $\langle u_1, u_2 \rangle = \gamma$ and the generated data point is either $X = u_1$ or $X = u_2$.*
2. *Correlated coefficients: $\mathbb{P}[s_1 = s_2 = 1] = \gamma$ and the modes are orthogonal, $\langle u_1, u_2 \rangle = 0$.*

We now present a more technical version of Assumption 2.

Assumption 4 (p_z structure)**.** *Let $z \sim p_z$. We assume that for $i, j \in [m_G]$,*

$$\Pr[z_i = 1] = \Theta\left(\frac{1}{m_G}\right), \ \Pr[z_i = z_j = 1] \quad = \frac{1}{m_G^2 \mathrm{polylog}(d)} \tag{12}$$

We set $\Pr[z_i = z_j = 1] = \frac{1}{m_G^2 \mathrm{polylog}(d)}$ to ensure that that the output of the generator is only made of one mode with probability $1 - o(1)$.

In the proof, we actually consider a more complicated version of the discriminator

$$D_{\mathcal{W}}(X) = \mathrm{sigmoid}\left(a \sum_{i \in [m_D]} \sigma(\langle w_i, X \rangle) + \lambda b\right), \text{ where } \sigma(z) \quad = \begin{cases} z^3 & \text{if } |z| \leq \Lambda \\ 3\Lambda^2 z - 2\Lambda^3 & \text{if } z > \Lambda \\ 3\Lambda^2 z + 2\Lambda^3 & \text{otherwise} \end{cases}, \tag{13}$$

where $\Lambda = d^{0.2}$. $\sigma(\cdot)$ is the truncated degree-3 activation function—it is thus made Lipschitz, which is only needed in the proof to *deal* with the case where the generator is trained much faster than the discriminator. Note that this latter case is uncommon in practice.

We now present the technical version of Parametrization 3.1.

Parametrization B.1. *When running SGDA and nSGDA on GAN, we set the parameters as*

*– **Initialization**: $b^{(0)} = 0, \ a^{(0)} \sim \mathcal{N}\left(0, \frac{1}{m_D \mathrm{polylog}(d)}\right), \ w_i^{(0)} \sim \mathcal{N}\left(0, \frac{1}{d}\mathbf{I}\right), \ v_j^{(0)} \sim \mathcal{N}\left(0, \frac{1}{d^2}\mathbf{I}\right)$ for $i \in [m_D]$, $j \in [m_G]$.*

*– **Number of iterations**: we run SGDA for $t \leq T_0$ iterations where T_0 is the first iteration such that $\|\nabla \mathbb{E}[L_{\mathcal{V}^{(T_0)}, \mathcal{W}^{(T_0)}}(X, z)]\|_2 \leq 1/\mathrm{poly}(d)$. . For nSGDA, we run for $T_1 = \tilde{\Theta}\left(\frac{1}{\eta_D}\right)$ iterations.*

*– **Step-sizes**: For SGDA, $\eta_D, \eta_G \in (0, \frac{1}{\mathrm{poly}(d)})$. For nSGDA, $\eta_D \in (0, \frac{1}{\mathrm{poly}(d)}]$, $\eta_G = \frac{\eta_D}{\mathrm{polylog}(d)}$.*

*– **Over-parametrization**: For SGDA, $m_D, m_G = \mathrm{polylog}(d)$ are arbitrarily chosen i.e. m_D may be larger than m_G or the opposite. For nSGDA, we set $m_D = \log(d)$ and $m_G = \log(d) \log\log d$.*

Regarding initialization, the discriminator's weights are sampled from a standard normal and its bias is set to zero. The weights of the generator are initialized from a normal with variance $1/d^2$ (instead of the $1/d$ in standard normal). Such a choice is explained as follows. In practice, the target $X \sim p_{data}$ is a 1D image, thus has entries in $[0, 1]^d$ and norm $O(\sqrt{d})$. Yet, we sample the initial generator's weights from $\mathcal{N}(0, \mathbf{I}_d/d)$ in this case. In our case, since $\|u_i\|_2 = 1$, the target $X = s_1 u_1 + s_2 u_2$ has norm $O(1)$. Therefore, we scale down the variance in the normal distribution by a factor of $1/d$ to match the configuration encountered in practice. Therefore, we also set $\lambda = \frac{1}{\sqrt{d}\mathrm{polylog}(d)}$ in (13) to ensure that the weights and the bias in the discriminator learn at the same speed.

Remark: In our theory, we consider the global version of nSGDA; $\|\mathbf{g}_{\mathcal{W}}^{(t)}\|_2$ in the update refers to $\|\mathbf{g}_{\mathcal{W}}^{(t)}\|_2 = \|\mathbf{g}_a^{(t)}\|_2 + \|\mathbf{g}_b^{(t)}\|_2 + \|\mathbf{g}_W^{(t)}\|_2$, where $\mathbf{g}_a^{(t)}$ is the stochastic gradient with respect to a, $\mathbf{g}_b^{(t)}$ with respect to b and $\mathbf{g}_W^{(t)}$ with respect to W.

B.2 Main results

We state now the technical version of Theorem 3.1 and Theorem 3.2.

Theorem B.1 (SGDA suffers from mode collapse)**.** *Let T_0, η_G, η_D and the initialization as defined in Parametrization B.1. Let t be such that $t \leq T_0$. Run SGDA for t iterations with step-sizes η_G, η_D. Then, with probability at least $1 - o(1)$, for all $z \in \{0,1\}^{m_G}$, we have:*

$$G_{\mathcal{V}}^{(t)}(z) = \alpha^{(t)}(z)(u_1 + u_2) + \xi^{(t)}(z),$$

where $\alpha^{(t)}(z) \in \mathbb{R}$ and $\xi^{(t)}(z) \in \mathbb{R}^d$ and for all $\ell \in [2]$, $|\langle \xi^{(t)}(z), u_\ell \rangle| = o(1)\|\xi^{(t)}(z)\|_2$ for every $z \in \{0,1\}^{m_G}$.

In the specific case where $\eta_D = \frac{\sqrt{d}\eta_G}{\mathrm{polylog}(d)}$, the model mode collapses i.e. $\|\xi^{(T_0)}(z)\|_2 = o(\alpha^{(T_0)}(z))$.

Theorem 3.1 indicates that with SGDA and any step-size configuration, the generator either does not learn the modes at all – when $\alpha^{(t)}(z) = 0,\ G_{\mathcal{V}}^{(t)}(z) = \xi^{(t)}(z)$ – or learns an average of the modes – when $\alpha^{(t)}(z) \neq 0,\ G_{\mathcal{V}}^{(t)}(z) \approx \alpha^{(t)}(z)(u_1 + u_2)$.

Theorem B.2 (nSGDA recovers modes separately)**.** *Let T_1, η_G, η_D and the initialization as defined in Parametrization B.1. Run nSGDA for T_1 iterations with step-sizes η_G, η_D. Then, the generator learns both modes u_1, u_2 i.e., for $\ell \in \{1,2\}$*

$$\Pr_{z \sim p_z}\left(\Big\| \frac{G_{\mathcal{V}}^{(T_1)}(z)}{\|G_{\mathcal{V}}^{(T_1)}(z)\|_2} - u_\ell \Big\|_2 = o(1) \right) = \tilde{\Omega}(1).$$

C Notations

Let us also write $\tau_b = \lambda$ as the scaling factor of the bias. We can easily observe that at every step, all of $w_i^{(t)}$ and $v_i^{(t)}$ lies in the span of $\{w_j^{(0)}, v_j^{(0)}, u_1, u_2\}$. Therefore, let us denote

$$w_i^{(t)} = \sum_{j \in [m_D]} \alpha(w_i, w_j, t) \frac{w_j^{(0)}}{\|w_j^{(0)}\|_2} + \sum_{j \in [m_G]} \alpha(w_i, v_j, t) \frac{v_j^{(0)}}{\|v_j^{(0)}\|_2} + \sum_{j \in [2]} \alpha(w_i, u_j, t) \frac{u_j}{\|u_j\|}$$

and $v_i^{(t)}$ as $\alpha(v_i, *, t)$, where $\alpha(*, *, *) \in \mathbb{R}$.

Let us denote

$$f(X) = a \left(\sum_{i \in [m_D]} \sigma(\langle w_i, X \rangle) \right) + \tau_b b$$

as the function in discriminator without going through sigmoid, and define $h(X) = \sum_{i \in [m_D]} \sigma(\langle w_i, X \rangle)$.

Gradient The gradient of $L(X, z)$ is given as:

$$\nabla_a L(X,z) = -\,\mathrm{Sigmoid}(-f(X))h(X) + \mathrm{Sigmoid}(f(G(z)))h(G(z))$$
$$\nabla_b L(X,z) = -\,\mathrm{Sigmoid}(-f(X)) + \mathrm{Sigmoid}(f(G(z)))$$
$$\nabla_{w_i} L(X,z) = -\,\mathrm{Sigmoid}(-f(X))a\sigma'(\langle w_i, X \rangle)X + \mathrm{Sigmoid}(f(G(z)))a\sigma'(\langle w_i, G(z) \rangle)G(z)$$
$$\nabla_{v_i} L(X,z) = -1_{z_i=1}\,\mathrm{Sigmoid}(f(G(z)))a \sum_{j \in [m_D]} \sigma'(\langle w_i, G(z) \rangle)w_i$$

We use $a^{(t)}, b^{(t)}, w_i^{(t)}, v_i^{(t)}$ to denote the value of those weights at iteration t.

We use $a = b \pm c$ for $c \in \mathbb{R}^*$ to denote: (1). $a \in [b - c, b + c]$ if $a, b \in \mathbb{R}$, (2). $\|a - b\|_2 \le c$ if a, b are vectors.

For simplicity, we focus on the case when all $\Pr[z_i = 1]$ are equal. The other cases can be proved similarly (by replacing the $1/m_G$ factor in the generators update by the exact value of $\Pr[z_i = 1]$).

D INITIALIZATION CONDITIONS AND THREE REGIME OF LEARNING

We first show the following Lemma regarding initialization:

Let

$$A_{i,\ell} = \frac{1}{2}\sigma'(\langle w_i^{(0)}, u_\ell\rangle)\operatorname{sign}(\langle w_i^{(0)}, u_\ell\rangle)$$

and

$$B_{i,j} = \frac{1}{m_G}\sigma'(\langle w_i^{(0)}, v_j^{(0)}\rangle)\operatorname{sign}(\langle w_i^{(0)}, v_j^{(0)}\rangle)$$

and

$$C_{i,\ell} = \sigma'(\langle v_i^{(0)}, u_\ell\rangle)$$

Let $A = \max_{i\in[m_D],\ell\in[2]} A_{i,\ell}$, $B = \max_{i\in[m_D],j\in[m_G]} B_{i,j}$, $C = \max_{i\in[m_G],\ell\in[2]} C_{i,\ell}$, we have: Using a corollary of Proposition G.1 in Allen-Zhu & Li (2020):

Lemma D.1. *For every $\eta_D, \eta_G > 0$, we have that: with probability at least $1 - o(1)$, we have that: $A = \frac{\text{polyloglog}(d)}{\sqrt{d}}$, $B = \frac{\text{polyloglog}(d)}{d m_G}$. Moreover, with probability at least $1 - o(1)$, one and only one of the following holds:*

1. *(Discriminator trains too fast): $\eta_G B < \frac{1}{\text{polylog}(d)}\eta_D A$;*
2. *(Balanced discriminator and generator): $\eta_G B > \frac{1}{\text{polylog}(d)}\eta_D A$, $\eta_D A > \eta_G B(1 + \frac{1}{\text{polyloglog}(d)})$;*
3. *(Generator trains too fast): $\eta_D A < \eta_G B(1 - \frac{1}{\text{polyloglog}(d)})$.*

This Lemma implies that in case 2, $\eta_G = \tilde{\Theta}(\sqrt{d})\eta_B$.

We will show the following induction hypothesis for each case. Intuitively, in case one we have the following learning process: (too powerful D).

1. At first D starts to learn, then because of the learning rate of G is too small, so D just saturate the loss to make the gradient to zero.

In case two we have: ("balanced" D and G but still not enough).

1. At first D starts to learn one u_j in each of the neuron.
2. However, the generator still could not catch up immediate after D learns one u_j, so D starts to a mixture of u_1, u_2 in its neurons since u_1, u_2 are positively correlated.
3. After that G starts to learn, however since D already stuck at the mixtures of u_1, u_2, so G is only able to learn mixtures of u_1, u_2 as well.

In case three we have: (Too powerful G)

1. G starts to learn without D learning any meaningful signal yet, so G aligns its outputs with the (close to random) weights of D and just pushes the discriminator to zero. In this case, G simply learns something random to fool D instead of learning the signals.

Moreover, similar to Lemma D.1, we also have the following condition regarding the gap between the top one and the second largest one in terms of correlation:

Lemma D.2. *Let*

$$i_D, \ell_D = \underset{i\in[m_D], \ell\in[2]}{\arg\max} \; A_{i,\ell}$$

Let

$$i_G, j_G = \underset{i\in[m_D], j\in[m_G]}{\arg\max} \; B_{i,j}$$

Then with probability at least $1 - o(1)$ *over the random initialization, the following holds:*

$$\forall i, \ell \neq i_D, \ell_D : A_{i_D,\ell_D} \geq A_{i,\ell}\left(1 + \frac{1}{\text{polyloglog}(d)}\right)$$
$$\forall i, j \neq i_G, j_G : B_{i_G,j_G} \geq B_{i,j}\left(1 + \frac{1}{\text{polyloglog}(d)}\right)$$

and

$$A = \frac{\Theta(\sqrt{\log\log(d)})}{\sqrt{d}}, \quad B, C = \frac{\Theta(\sqrt{\log\log(d)})}{d}$$

For simplicity, we also define $i^* = i_D$.

E Critical Lemma

The proof heavily relies on the following Lemma about tensor power method, which is a corollary of Lemma C.19 in Allen-Zhu & Li (2020).

Lemma E.1. *For every* $\delta \in (0, 0.1)$, *every* $C > 10$, *for every sequence of* $x_t, y_t > 0$ *such that* $x_0 > (1 + 10\delta)y_0$, *suppose there is a sequence of* $S_t \in [0, C]$ *such that for* $\eta \in \left(0, \frac{1}{\text{poly}(C/\delta)}\right)$*:*

$$x_{t+1} \geq x_t + \eta S_t x_t^2$$
$$y_t \leq y_t + \eta S_t(1 + \delta)y_t^2$$

For every $\tau > 0$, *let* T_0 *be the first iteration where* $x_t > \tau$, *then we must have:*

$$y_{T_0} \leq \frac{y_0}{\text{poly}(\delta)}$$

Moreover, if all $S_t \geq H$ *for some* $H > 0$, *then* $T_0 \leq O\left(\frac{1}{\eta H x_0}\right)$.

Similar to the Lemma above, one can easily show the following auxiliary Lemma:

Lemma E.2. *Suppose there are sequences* $a_t, b_t \in \mathbb{R}^d$ *such that* $a_0, b_0 > 0$ *with* $a_0 < 0.82b_0$. *Suppose there exists a sequence of* $C_t \in (0, d)$ *such that*

$$a_{t+1} \leq a_t - \eta_D C_t b_t$$
$$b_{t+1} \geq b_t - 1.0000001\eta_D C_t a_t$$

Then we must have that for every $t \leq T$ *where* T *is the first iteration such that* $a_T \leq 0$, *then the following holds:*

$$a_t = a_0 - \Theta\left(\eta \sum_{s\leq t-1} C_s\right)$$
$$\sum_{s\leq t} |a_t C_t \eta_D| \leq 0.49b_0$$

Moreover, if in addition that $a_0 < \frac{1}{C}b_0$ *for any* $C > 100$, *then we must have:*

$$\sum_{s\leq t} |a_t C_t \eta_D| \leq \frac{10}{C}b_0$$

In the end, we have the following comparison Lemma, whose proof is obvious:

Lemma E.3. *Suppose $a_t, b_t > 0$ satisfies that $a_0, b_0 \le 1$, and the update of a_t, b_t is given as: For some values $C > 0$ and $C_t \in [0, \mathrm{poly}(d)]$:*

$$a_{t+1} = a_t + \eta_D C_t \tag{14}$$

$$b_{t+1} = b_t + \eta_D \left[\frac{1}{C}, 1\right] \times C_t \tag{15}$$

Then let T be the first iteration where $a_T \ge 2C$, we must have:

$$b_T \in [1, 2C + 1]$$

Using this Lemma, we can directly prove the following Lemma:

Lemma E.4. *For every $\eta_D, \eta_G \in \left(0, \frac{1}{\mathrm{poly}(d)}\right]$ such that $\eta_G = \eta_D \Gamma$ for $\Gamma = \tilde{\Theta}(\sqrt{d})$, suppose there are vectors $p_t, q_{i,t} \in \mathbb{R}^d$ ($i \in [m_G]$) and a value $a_t \in \mathbb{R}, H > 0$ satisfies that for a sequence of $H_{i,t} \in [H, 1]$ for $i \in [m_G]$, $G_t = \tilde{\Theta}(\sum_{i \in [m_G]} H_{i,t})$, a value $\tau = \tilde{O}(d^{-0.5})$, and a vector $\beta_t \in$ span$\{u_1, u_2\}$ with $\|\beta_t\|_2 = O(1)$: For all $i \in [m_G]$ and $t \ge 0$:*

$$\|q_{i,0}\|_2 = \tilde{\Theta}(d^{-0.49}), \|p_0\|_2 = \log^{\Theta(1)}(d), 0 < a_0 \le 0.819\|p_0\|_2 \quad \frac{\langle q_{i,0}, p_0 \rangle}{\|q_{i,0}\|_2, \|p_{i,0}\|_2} \ge 1 - o(1)$$

$$p_t = p_t - \eta_D \sum_{i \in [m_G]} G_{i,t} a_t \sigma'(\langle p_t, q_{i,t} \rangle) q_{i,t} + \tilde{O}(\eta_D a_t \gamma_t) \beta_t$$

$$a_t = a_t - \eta_D \sum_{i \in [m_G]} G_{i,t} \sigma(\langle p_t, q_{i,t} \rangle) \pm \tilde{O}(\eta_D \gamma_t)$$

$$q_{i,t} = \left(q_{i,t} + \eta_G H_{i,t} a_t \left(\sigma'(\langle p_t, q_{i,t} \rangle) + \sum_{j \in [m_G]} \gamma_{i,j,t} \sigma'(\langle p_t, q_{j,t} \rangle) \right) p_t \pm \eta_G |a_t| \tilde{O}\left(\tau \|q_{i,t}\|_2\right)^2 \right)$$

In addition, we have: $\gamma_{i,j,t} = \tilde{O}(1)$, and

$$\max_{i \in [m_G]} \|q_{i,t}\|_2 \in \left(0, \frac{1}{\mathrm{polylog}(d)}\right] \cup [\mathrm{polylog}(d), +\infty) \implies \forall i, j \in [m_G], H_{i,t} = \tilde{\Theta}(G_t), \gamma_{i,j,t} = \tilde{\Theta}(1)$$

Then we must have that: let T be the first iteration where $a_T \le 0$, we have: for every $t \le T$: there is a scaling factor $\ell_t = \Theta(1)$ such that

$$\|p_t - \ell_t p_0\|_2 \le o(1)\|p_0\|_2, \quad \|\Pi_{\mathit{span}\{u_1, u_2, p_0\}^\perp}(p_t - p_0)\|_2 \le d^{-0.6}\|p_0\|_2$$

Moreover, for every $i, j \in [m_G]$, $\|q_{i,t}\|_2 = \tilde{\Theta}(\|q_{j,t}\|_2)$ and $\|q_{i,T}\|_2 \ge \tilde{\Theta}(\sqrt{\Gamma})$, and as long as $\max_{i \in [m_G]} \|q_{i,t}\|_2 \ge \mathrm{polylog}(d)$, we have that $a_t \|q_{i,t}\|_2 \ge \mathrm{polylog}(d)$.

Moreover,

$$\|\Pi_{\mathit{span}\{u_1, u_2, p_0\}^\perp}(q_{i,t} - q_{i,0})\|_2 \le d^{-0.6}\|q_{i,t}\|_2$$

proof of Lemma E.4. For simplicity we consider the case when $H = \tilde{\Omega}(1)$, the other cases follow similarly.

To proof this result, we maintain the following decomposition of p_t and $q_{i,t}$ as:

$$p_t = \alpha(t) p_0 + \beta(t) + \gamma(t)$$

Where $\beta(t) \in$ span$\{u_1, u_2\}$ and $\gamma(t) \perp$span$\{u_1, u_2, p_0\}$. Note that $\alpha(0) = 1, \beta(0) = \gamma(0) = 0$.

$$q_{i,t} = \alpha(i,t) p_0 + \beta(i,t) + \gamma(i,t)$$

Where $\beta(i,t) \in \text{span} u_1, u_2$ and $\gamma(i,t) \perp \text{span}\{p_0, u_1, u_2\}$.

We maintain the following induction hypothesis (which we will prove at the end): For some $\mu = 0.00001$ and $C_1 = d^{-0.1}, C_2 = d^{-0.6}$, we have:

1. Through out the iterations, $\alpha(t) \geq 0.5$ and $\|\beta(t)\|_2 \leq 0.5(1-\mu)C_1, \|\gamma(t)\|_2 \leq 0.5(1-\mu)C_2$.
2. $\alpha(i,t) \in (0, \tilde{O}\sqrt{\Gamma})$ and $\|\beta(i,t)\|_2 \leq C_1\alpha(i,t) + \|\beta(i,0)\|_2, \quad \|\gamma(i,t)\|_2 \leq C_2\alpha(i,t) + \|\gamma(i,0)\|_2$

The induction hypothesis implies that through out the iterations, $\langle q_{i,t}, p_t \rangle = \tilde{\Omega}(\|q_{i,t}\|_2)$.

We can now write down the update of $a_t, \alpha's, \beta's$ and $\gamma's$ as:

$$a_{t+1} = a_t - \eta_D \left(\sum_{i \in [m_G]} G_{i,t} \sigma(\langle p_t, q_{i,t} \rangle) \pm \tilde{O}(1) \right) \tag{16}$$

$$\alpha(t+1) = \alpha(t) - \eta_D a_t \sum_{i \in [m_G]} G_{i,t} \sigma'(\langle p_t, q_{i,t} \rangle) \alpha(i,t) \tag{17}$$

$$\beta(t+1) = \beta(t) - \eta_D a_t \sum_{i \in [m_G]} G_{i,t} \sigma'(\langle p_t, q_{i,t} \rangle) \beta(i,t) \pm \tilde{O}(\eta_D |a_t|) \tag{18}$$

$$\gamma(t+1) = \gamma(t) - \eta_D a_t \sum_{i \in [m_G]} G_{i,t} \sigma'(\langle p_t, q_{i,t} \rangle) \gamma(i,t) \tag{19}$$

By the induction hypothesis, we know that

$$\sigma'(\langle p_t, q_{j,t} \rangle) \geq \tilde{\Omega}\left(\|q_{j,t}\|_2^2 \times \frac{\Lambda^2}{\Gamma} \right)$$

Moreover, we have that let $h_{i,t} := \left(\sigma'(\langle p_t, q_{i,t} \rangle) + \sum_{j \in [m_G]} \tilde{\Theta}(\sigma'(\langle p_t, q_{j,t} \rangle)) \right)$

$$\alpha(i,t+1) = \left(\alpha(i,t) + \eta_G H_{i,t} a_t h_{i,t} (1 \pm \tilde{O}(\tau^2 \Lambda^2 / \Gamma) \alpha(t) \right) \tag{20}$$

$$\beta(i,t+1) = \left(\beta(i,t) + \eta_G H_{i,t} a_t h_{i,t} \left(\beta(t) \pm \tilde{O}(\tau^2 \Lambda^2 / \Gamma) \right) \right) \tag{21}$$

$$\gamma(i,t+1) = \left(\alpha(i,t) + \eta_G H_{i,t} a_t h_{i,t} \left(\gamma(t) \pm \tilde{O}(\tau^2 \Lambda^2 / \Gamma) \right) \right) \tag{22}$$

From these formula, we can easily that as long as (1). $\alpha(t) \geq 0.5$ and $\|\beta(t)\|_2 \leq 0.5(1-\mu)C_1, \|\gamma(t)\|_2 \leq 0.5(1-\mu)C_2$, (2). $C_1, C_2 = \tilde{\Omega}(\tau^2 \Lambda^2 / \Gamma)$, we must have that $\alpha(i,t) > 0$ and $\|\beta(i,t)\|_2 \leq C_1\alpha(i,t) + \|\beta(i,0)\|_2, \quad \|\gamma(i,t)\|_2 \leq C_2\alpha(i,t) + \|\gamma(i,0)\|_2$. Therefore, it remains to only prove (1) in the induction hypothesis. Moreover, it is easy to observe that $\alpha(i,t) = \tilde{\Theta}(\alpha(j,t))$ for all $i, j \in [m_G]$ and all t.

Now, we divide the update process into two stages:

Before all $\|q_{i,t}\|_2 = \Omega(\Lambda)$. Let's call these iterations $[T_1]$ Let us consider $T_{i,1}$ such that for all $t \in [T_{i,1}]$ when $q_{i,t} = O(\Lambda)$ and $a_t = \Omega(1)$. In these iterations, by the update rule, we have

$$q_{i,t} = q_{i,t} + \tilde{\Omega}(\eta_G)\sigma'(\langle p_t, q_{i,t}\rangle)p_t \pm \tilde{O}(\eta_G \tau^2 \|q_{i,t}\|_2^2)$$

By the induction hypothesis, we can simplify the update as:

$$\langle q_{i,t}, p_0\rangle \geq \langle q_{i,t}, p_0\rangle + \tilde{\Omega}\left(\eta_G \sigma'(\langle q_{i,t}, p_0\rangle)\right)$$

Therefore, we know that $T_{i,1} \leq \tilde{O}\left(\frac{d^{0.49}}{\eta_G}\right)$ and

$$\sum_{t \leq T_{i,1}} \sigma'(\langle q_{i,t}, p_0\rangle), \sum_{t \leq T_{i,1}} \sigma'(\langle q_{i,t}, p_t\rangle) \leq \tilde{O}\left(\frac{\Lambda}{\eta_G}\right) \tag{23}$$

Together with the induction hypothesis, the fact that $\alpha(i,t) = \tilde{\Theta}(\alpha(j,t))$, the fact that $\sigma(\langle p_t, q_{i,t}\rangle) = \tilde{\Theta}(\sigma'(\langle p_t, q_{i,t}\rangle)\|q_{i,t}\|_2)$ and update formula Eq equation 16 equation 31 equation 18 equation 19, we know that for all $t \leq \max\{T_{i,1}\}$:

$$a_t = a_0 \pm \tilde{O}\left(\frac{\eta_D \Lambda^2}{\eta_G}\right) = a_0 \pm \tilde{O}(d^{-0.01}) \tag{24}$$

$$\alpha(t) = \alpha(0) \pm \tilde{O}\left(\frac{\eta_D \Lambda^2}{\eta_G}\right) = \alpha(0) \pm \tilde{O}(d^{-0.01}) \tag{25}$$

$$\|\beta(t)\|_2 \leq \tilde{O}\left(\frac{\eta_D \Lambda^2}{\eta_G}\right) C_1 + \tilde{O}\left(\frac{\eta_D \|\beta(i,0)\|_2 \Lambda}{\eta_G}\right) \leq \tilde{O}(d^{-0.01})C_1 \tag{26}$$

$$\|\gamma(t)\|_2 \leq \tilde{O}\left(\frac{\eta_D \Lambda^2}{\eta_G}\right) C_2 + \tilde{O}\left(\frac{\eta_D \|\gamma(i,0)\|_2 \Lambda}{\eta_G}\right) \leq \tilde{O}(d^{-0.01})C_2 \tag{27}$$

When all $\|q_{i,t}\|_2 = \Omega(\Lambda)$: In this case, since $\|p_0\|_2 = \omega(1)$, we know that $\langle p_t, q_{i,t}\rangle = \omega(\Lambda)$, so $\sigma(\langle p_t, q_{i,t}\rangle)$ acts on the linear regime, which means that:

$$\sigma(\langle p_t, q_{i,t}\rangle) = (1 \pm o(1))3\Lambda^2\langle p_t, q_{i,t}\rangle, \quad \sigma'(\langle p_t, q_{i,t}\rangle) = (1 \pm o(1))3\Lambda^2$$

Therefore, we know that:

$$a_{t+1} \leq a_t - (1 - o(1))\eta_D \left(\sum_{i \in [m_G]} G_{i,t} 3\Lambda^2 \|q_{i,t}\|_2\right) \alpha(t)\|p_0\|_2 \tag{28}$$

$$\alpha(t+1)\|p_0\|_2 \geq \alpha(t)\|p_0\|_2 - (1 + o(1))\eta_D \left(\sum_{i \in [m_G]} G_{i,t} 3\Lambda^2 \|q_{i,t}\|_2\right) a_t \tag{29}$$

Now, using the fact that $a_0 \leq 0.819\alpha(0)$ and with Eq equation 24 and Eq equation 25, apply Lemma E.2 we have that until $a_t \leq 0$,

$$\sum_t \eta_D \left(\sum_{i \in [m_G]} G_{i,t} 3\Lambda^2 \|q_{i,t}\|_2\right) a_t \leq 0.49\|p_0\|_2 \tag{30}$$

Plug in to the update rule:

$$\alpha(t+1) = \alpha(t) \pm (1 + o(1))\eta_D a_t \sum_{i \in [m_G]} G_{i,t} 3\Lambda^2 \alpha(i,t) \tag{31}$$

$$= \alpha(t) \pm (1 + o(1))\eta_D a_t \sum_{i \in [m_G]} G_{i,t} 3\Lambda^2 \frac{\|q_{i,t}\|_2}{\|p_0\|} \tag{32}$$

$$\|\beta(t+1)\|_2 \le \|\beta(t)\|_2 + (1+o(1))\eta_D a_t \left(\sum_{i\in[m_G]} G_{i,t} 3\Lambda^2 \beta(i,t) + \tilde{O}(1) \right) \tag{33}$$

$$\le \|\beta(t)\|_2 + \eta_D(1+o(1))a_t \left(\sum_{i\in[m_G]} G_{i,t} 3\Lambda^2 \frac{\|q_{i,t}\|_2 C_1}{\|p_0\|_2} + \tilde{O}(1) \right) \tag{34}$$

$$\le \|\beta(t)\|_2 + \eta_D(1+o(1))a_t \left(\sum_{i\in[m_G]} G_{i,t} 3\Lambda^2 \frac{\|q_{i,t}\|_2 C_1}{\|p_0\|_2} \right) \tag{35}$$

$$\|\gamma(t+1)\|_2 \le \|\gamma(t)\|_2 + \eta_D(1+o(1))a_t \sum_{i\in[m_G]} G_{i,t} 3\Lambda^2 \gamma(i,t) \tag{36}$$

$$\le \|\gamma(t)\|_2 + \eta_D(1+o(1))a_t \sum_{i\in[m_G]} G_{i,t} 3\Lambda^2 \frac{\|q_{i,t}\|_2 C_2}{\|p_0\|_2} \tag{37}$$

We directly complete the proof of the induction hypothesis using Eq equation 30.

Now it remains to prove that $\|q_{i,T}\|_2 = \Omega(\sqrt{\Gamma})$. Compare the update rule of $q_{i,t}$ and a_t we have:

$$a_{t+1} = a_t - \tilde{\Theta}(\eta_D) G_t \left(\sum_{i\in[m_G]} \Lambda^2 \|q_{i,t}\|_2 \right) \tag{38}$$

and

$$\sum_{i\in[m_G]} \|q_{i,t+1}\|_2 = \|q_{i,t}\|_2 + \tilde{\Theta}(\eta_G) G_t \Lambda^2 a_t \tag{39}$$

We can directly conclude that $\|q_{i,t}\|_2 \le \tilde{O}\left(\sqrt{\Gamma}\right)$ and $\|q_{i,T}\|_2 = \tilde{\Omega}(\sqrt{\Gamma})$.

□

Lemma E.5. *For every $\eta_D, \eta_G \in \left(0, \frac{1}{\mathrm{poly}(d)}\right]$ such that $\eta_G = \eta_D \Gamma$ for $\Gamma \ge \tilde{\Omega}(\sqrt{d})$, suppose for sufficiently large $C = \mathrm{poly}(\log(d) m_D)$ there are vectors $\{q_{i,t}\}_{i\in[m_G]}$, $\{p_i\}_{i\in[m_D]}$ in $\mathbb{R}^d$ such that $\|p_i\|_2 = 1, \langle p_i, p_{i'} \rangle \le \tilde{O}(1/\sqrt{d})$ for i, i', values $H_{i,j,t}, G_{i,t} \in \left[\frac{1}{C^2}, C^2\right]$ and a value $a_0 \ge 0$ satisfies that:*

$$a_0 = \frac{1}{\mathrm{polylog}(d)}, \|q_{j,0}\|_2 = \tilde{\Theta}(\Lambda); \quad q_{j,0} = \sum_{i\in[m_D]} a_i p_i + \xi_j, a_i \ge 0, \|\xi_i\|_2 \le \frac{1}{C}\|q_{j,0}\|_2 \tag{40}$$

$$a_{t+1} = a_t - \eta_D \left(H_{i,j,t} \sum_{i\in[m_D], j\in[m_G]} \sigma(\langle p_i, q_{j,t} \rangle) \right) \tag{41}$$

$$q_{i,t+1} = q_{i,t} + \eta_G a_t G_{i,t} \sum_{j\in[m_D]} \left(\sigma'(\langle p_j, q_{i,t} \rangle) \left(p_j \pm \frac{1}{C} \right) \right) \tag{42}$$

Then we must have: within $T = \tilde{O}\left(\frac{\sqrt{\Gamma}}{\eta_G}\right)$ many iterations, we must have that $a_t \le 0$ and $\max_{j\in[m_G]} \|q_{j,T}\|_2 = \tilde{\Theta}(\sqrt{\Gamma})$. Moreover, for every $t \le T$, we have: for every $j \in [m_G]$,

$$\sum_{i\in[m_D]} \sigma(\langle p_i, q_{j,t} \rangle) = \Omega\left(\max_{i\in[m_D]} \sigma'(\langle p_i, q_{j,t} \rangle) \|q_{j,t}\|_2 \right)$$

and

$$\max_{i\in[m_D]} \langle p_i, q_{j,t}\rangle \geq \left(1 - \frac{1}{C^{0.2}}\right)\|q_{j,t}\|_2$$

Proof of Lemma E.5. Let us denote $r_{i,t} = \max_{j\in[m_D]}\{\langle p_j, q_{i,t}\rangle\}$.

By the update rule, we can easily conclude that:

$$r_{i,t+1} = r_{i,t} + \eta_G G_{i,t}\left(1 - \frac{1}{C^{0.5}}\right)\sigma'(r_{i,t})$$

On the other hand, let us write $q_{i,t} = \sum_{j\in[m_D]} \alpha_{i,j,t} q_j + \xi_{i,t}$, where $\alpha_{i,j,t} \geq 0$. We know that:

$$\|\xi_{i,t+1}\|_2 \leq \|\xi_{i,t}\|_2 + \eta_G G_{i,t}\frac{m_D}{C}\sigma'(r_{i,t}) \tag{43}$$

By the comparison Lemma E.3 we can easily conclude that for every t,

$$\|\xi_{i,t}\|_2 \leq \frac{1}{C^{0.5}} r_{i,t}$$

This implies that: there exists values $u_t \in [1/C^2, C^2]$ such that

$$a_{t+1} = a_t - \eta_D u_t \sum_{i\in[m_G]} \sigma(r_{i,t}) \tag{44}$$

Comparing this with the update rule of $r_{i,t}$, we know that for every t with $a_t \geq 0$, we must have:

$$r_{i,t} = \tilde{O}\left(\sqrt{\Gamma}\right), \quad r_{i,T} = \tilde{\Theta}(\sqrt{\Gamma})$$

□

Lemma E.6 (Auxiliary Lemma)**.**

For every $g > 0$ we must have: Sigmoid$(-gx - b)x$ *is a decreasing function of x as long as $gx > 1$ and $gx + b > 0$.*

Lemma E.7. *For $a_t, b_t, c_t, d_t \in \mathbb{R}^d$ be defined as: $a_0, c_0, d_0 = \frac{1}{\text{polylog}(d)}$, $|b_t| \leq O(\log d)$ and $|b_t| \leq \min\{a_t c_t^3, a_t d_t^3\}$.*

$$a_{t+1} = a_t + \eta_D \frac{1}{2}\left(\left(1 \pm \frac{1}{\text{polylog}(d)}\right)\text{Sigmoid}(-a_t c_t^3 - b_t)c_t^3 + \left(1 \pm \frac{1}{\text{polylog}(d)}\right)\text{Sigmoid}(-a_t d_t^3 - b_t)d_t^3\right) \tag{45}$$

$$c_{t+1} = c_t + \eta_D \frac{3}{2}\left(\left(1 \pm \frac{1}{\text{polylog}(d)}\right)\text{Sigmoid}(-a_t c_t^3 - b_t)c_t^2 a_t\right) \tag{46}$$

$$d_{t+1} = d_t + \eta_D \frac{3}{2}\left(\left(1 \pm \frac{1}{\text{polylog}(d)}\right)\text{Sigmoid}(-a_t d_t^3 - b_t)d_t^2 a_t\right) \tag{47}$$

Then we have: for every $t \in \left(\frac{\text{polylog}(d)}{\eta_D}, \frac{\text{poly}(d)}{\eta_D}\right]$, we must have:

$$a_t = \sqrt{\frac{2}{3}}\left(1 \pm \frac{1}{\text{polylog}(d)}\right)c_t \tag{48}$$

$$c_t = \left(1 \pm \frac{1}{\text{polylog}(d)}\right)d_t \tag{49}$$

Proof of Lemma E.7. By the update formula, we can easily conclude that for $t \leq \frac{\text{poly}(d)}{\eta_D}$, we have that $a_t, c_t, d_t \in \left[\frac{1}{\text{polylog}(d)}, \text{polylog}(d)\right]$. This implies that for every $t \in \left[\frac{\text{polylog}(d)}{\eta_D}, \frac{\text{poly}(d)}{\eta_D}\right]$, we have that

$$a_t c_t^3, a_t d_t^3 \in [1, O(\log d)]$$

Apply Lemma E.6 we have that: As long as $a_t > 3(c_t + d_t)$, we must have that

$$\text{Sigmoid}(-a_t c_t^3 - b_t)c_t^3 + \text{Sigmoid}(-a_t d_t^3 - b_t)d_t^3 < \text{Sigmoid}(-a_t c_t^3 - b_t)c_t^2 a_t + \text{Sigmoid}(-a_t d_t^3 - b_t)d_t^2 a_t$$

This implies that

$$\frac{a_{t+1}}{3} - \frac{a_t}{3} < c_{t+1} + d_{t+1} - c_t - d_t$$

Note that initially, $a_0, c_0, d_0 = \frac{1}{\text{polylog}(d)}$. This implies that when $t \geq \frac{\text{polylog}(d)}{\eta_D}$, we must have that $a_t \leq 4(c_t + d_t)$, therefore $c_t + d_t = \Omega(1)$. Similarly, we can prove that $a_t \geq 0.1 \min\{c_t, d_t\}$.

as long as $c_t > d_t$, we must have:

$$\text{Sigmoid}(-a_t c_t^3 - b_t)c_t^2 a_t < \text{Sigmoid}(-a_t d_t^3 - b_t)d_t^2 a_t$$

Which implies that:

$$\frac{c_{t+1}}{1 + 1/\operatorname{polylog}(d)} - \frac{c_t}{1 + 1/\operatorname{polylog}(d)} < d_{t+1} - d_t \tag{50}$$

Note that initially, $c_0, d_0 = \frac{1}{\text{polylog}(d)}$ and when $t \geq \frac{\text{polylog}(d)}{\eta_D}$, $c_t + d_t = \Omega(1)$. This implies that for every $t \in \left[\frac{\text{polylog}(d)}{\eta_D}, \frac{\text{poly}(d)}{\eta_D}\right]$, we have: $c_t = \left(1 \pm \frac{1}{\text{polylog}(d)}\right) d_t$. Which also implies that $c_t, d_t \leq O(\log d)$.

Similarly, we can prove the bound for a_t.

□

F Induction hypothesis

F.1 Case 1: Balanced generator and discriminator

In this section we consider the case 2 in Lemma D.1. Here we give the induction hypothesis to prove the case of balanced generator and discriminator, this is the most difficult case and other cases are just simple modification of this one. Without loss of generality (by symmetry), let us assume that $a^{(0)} > 0$ and $a^{(0)} = \frac{1}{\text{polylog}(d)}$ (this happens with probability $1 - o(1)$).

We divide the training into five stages: For a sufficiently large $C = \text{polylog}(d)$

1. Stage 1: Before one of the $\alpha(w_i, u_j, t) \geq 1/C$. Call this exact iteration $T_{B,1}$.
2. Stage 2: After $T_{B,1}$, before $T_{B,2} = T_{B,1} + \frac{1}{\eta_D 2^{\sqrt{\log(d)}}}$.
3. Stage 3: After $T_{B,2}$, before one of the $\alpha(v_i, u_j, t) \geq d^{-0.49}$. Call this exact iteration $T_{B,3}$.
4. Stage 4: After $T_{B,3}$, before $a^{(t)} \leq \tilde{O}\left(\frac{1}{\Lambda^2 d^{1/4}}\right)$. Call this exact iteration $T_{B,4}$.
5. Stage 5: After $T_{B,4}$, until convergence.

We maintain the following things about α and a, b during each stage:

Stage 1 : We maintain: For every $t \leq T_{B,1}$:

1. (B.1.0). For all but the $i^* \in [m_D]$, and for all $j \in [m_G]$ (Below $*$ can be $w_{i'}, v_{j'}, u_\ell$ for every $i' \in [m_D], j' \in [m_G]$ and $\ell \in [2]$).

$$\forall * \neq u_1, u_2 : |\alpha(w_i, *, t) - \alpha(w_i, *, 0)| \leq \frac{1}{d^{0.9}}, \quad |\alpha(w_i, u_\ell, t) - \alpha(w_i, u_\ell, 0)| \leq \frac{C}{\sqrt{d}}$$

2. (B.1.1). For all $j \in [m_G]$:

$$|\alpha(v_j, *, t) - \alpha(v_j, *, 0)| \le \frac{\text{polyloglog}(d)}{d}$$

$$|\alpha(v_j, u_\ell, t)| \le \frac{1}{d}$$

3. (B.1.2). For $i^* = i$, we have that: for all $* \neq u_1, u_2$:

$$|\alpha(w_i, *, t) - \alpha(w_i, *, 0)| \le \frac{1}{d^{0.9}}$$

4. (B.1.3). a and b remains nice:

$$a^{(t)} \in (1 - 1/C, 1 + 1/C)a_0, |b^{(t)}| \le \frac{1}{C}$$

Stage 2 : For every $t \in [T_{B,1}, T_{B,2}]$.

1. (B.1.0), (B.1.1), (B.1.2) still holds.
2. (B.2.1): For $i = i^*$, we have:

$$a^{(t)}, \alpha(w_i, u_\ell, t) = \tilde{\Theta}(1)$$

Stage 3 : For every $t \in [T_{B,2}, T_{B.3}]$.

1. (B.1.0), (B.1.2) still holds.
2. (B.3.2): For every $j \in [m_G]$: For $* \neq w_{i^*}, u_1, u_2$, we have:

$$|\alpha(v_j, *, t) - \alpha(v_j, *, 0)| \le \frac{C^3}{\sqrt{d}} \|v_j^{(t)}\|_2$$

and

$$\alpha(v_i, u_\ell, t) \ge -O\left(\frac{1}{d}\right)$$

Moreover, let $\alpha(t) := \max_{j \in [m_G], \ell \in [2]} \langle v_j^{(t)}, u_\ell \rangle$, we have that:

$$|\alpha(v_j, w_{i^*}, t)| \le \text{polyloglog}(d)\alpha(t), \quad |\langle v_j^{(t)}, u_\ell \rangle| \le O(\alpha(t))$$

3. (B.3.3): Balanced update: for every X,

$$\text{Sigmoid}\left(-a^{(t)} \langle w_{i^*}^{(t)}, X \rangle^3 - b^{(t)}\right) \in \left[\frac{1}{\sqrt{d}\,\text{polylog}(d)}, \frac{1}{\text{polylog}(d)}\right] \text{Sigmoid}\left(b^{(t)}\right)$$

and

$$\text{Sigmoid}\left(-a^{(t)} \langle w_{i^*}^{(t)}, u_1 \rangle^3 - b^{(t)}\right) = \left(1 \pm \frac{1}{\text{polylog}(d)}\right) \text{Sigmoid}\left(-a^{(t)} \langle w_{i^*}^{(t)}, u_2 \rangle^3 - b^{(t)}\right)$$

Stage 4 : For every $t \in [T_{B,3}, T_{B.4}]$.

1. (B.3.1), (B.3.2) still holds.
2. (B.4.1) for $i = i^*$, we have that for all $* \neq u_1, u_2, w_i$:

$$|\alpha(w_i, *, t) - \alpha(w_i, *, 0)| \le \frac{C}{\sqrt{d}}$$

For all $* \in [u_1, u_2, w_i]$:

$$\alpha(w_i, *, t) = \Theta(\alpha(w_i, *, T_{B,3}))$$

3. For every $i, j \in [m_G]$: $\|v_i^{(t)}\|_2 = \tilde{\Theta}(\|v_j^{(t)}\|_2)$ and after $t = T_{B,4}$, we have that for every $i \in [m_G]$, $\|v_i^{(t)}\|_2 = \tilde{\Theta}(d^{1/4})$.
4. $|a^{(t)}|, |b^{(t)}| = O(\log(d))$.

Stage 5 : For every $t \in [T_{B,4}, T_0]$.

1. For every $i \in [m_D]$,
$$|\alpha(w_i, *, T_{B,4}) - \alpha(w_i, *, t)| \le d^{-0.1}$$
2. For every $i \in [m_G]$,
$$|\alpha(v_i, *, T_{B,4}) - \alpha(v_i, *, t)| \le d^{0.2}$$
3. $|a^{(t)}| \le \tilde{O}\left(\frac{1}{\Lambda^2 d^{1/4}}\right)$, and for every z:
$$\langle w_{i^*}^{(t)}, G^{(t)}(z)\rangle = \tilde{\Omega}(d^{1/4})$$

F.2 Case 2: Generator is dominating

We now consider another case where the generator's learning rate dominates that of the discriminator. This corresponds to case 3 in Lemma D.1. In this case, we divide the learning into four stages: For a sufficiently large $C = 2^{\sqrt{\log d}}$:

1. Before $\alpha(v_{j_G}, w_{i_G}, t) \ge d^{-0.49}$. Call this iteration $T_{G,1}$.
2. After $T_{G,1}$, before $\alpha(v_{j_G}, w_{i_G}, t) \ge \Lambda$. Call this iteration $T_{G,2}$.
3. After iteration $T_{G,2}$, before $a_t \le 0$. Call this iteration $T_{G,3}$.
4. After $T_{G,3}$.

We maintain the following induction hypothesis:

Stage 1 : In this stage, we maintain the following induction hypothesis: Let $\alpha(t) := \alpha(v_{j_G}, w_{i_G}, t)$, for every $t \le T_{G,1}$:

1. (G.1.1). For all $i \in [m_D]$, and for all $j \in [m_G]$:
$$|\alpha(w_i, *, t) - \alpha(w_i, *, 0)| \le \frac{C}{\sqrt{d}}$$
2. (G.1.2). For all $j \in [m_G]$, for all $* \neq w_{i_G}$:
$$|\alpha(v_j, *, t) - \alpha(v_j, *, 0)| \le \frac{C}{\sqrt{d}}\alpha(t)$$

Stage 2 : In this stage, we maintain: for every $t \in [T_{G,1}, T_{G,2}]$

1. (G.2.1). For every $i \in [m_D]$, we have:
$$|\alpha(w_i, *, t) - \alpha(w_i, *, 0)| \le \frac{1}{C}$$
2. (G.2.2). For every $j \in [m_G]$, $\alpha(v_i, w_{i_G}, t) \ge d^{-0.49}$.
3. For every $i \in [m_G]$, we have: for every $* \neq w_{i_G}$:
$$|\alpha(v_i, *, t) - \alpha(v_i, *, 0)| \le \frac{2}{C}|\alpha(v_i, w_{i_G}, t)|$$

Stage 3 : In this stage, we maintain: For every $t \in [T_{G,2}, T_{G,3}]$:

1. (G.2.1), (G.2.2) still holds.
2. For every $i \in [m_G]$, we have: for every $* = v_r$ or $* = u_\ell$:
$$|\alpha(v_i, *, t) - \alpha(v_i, *, 0)| \le \frac{2}{C}\|v_i^{(t)}\|_2$$

Stage 4 : In this stage, we maintain: For every $t \in [T_{G,3}, T_1]$:

1. (G.2.1) still holds.
2. For every $i \in [m_G]$, we have:
$$|\alpha(v_i, *, t) - \alpha(v_i, *, T_{G,3})| \leq \frac{1}{C} \|v_i^{(T_{G,3})}\|_2$$
3. $|\alpha_t| = \tilde{O}\left(\frac{1}{\Lambda^2 \sqrt{\eta_G/\eta_D}}\right)$, $\|v_i^{(t)}\|_2 = \tilde{\Theta}(\sqrt{\eta_G/\eta_D})$, and for all $z \neq 0$, $\sum_{i \in [m_D]} h(G^{(t)}(z)) = \tilde{\Theta}(\Lambda^2 \sqrt{\eta_G/\eta_D})$.

G Proof of the learning process in balanced case

For simplicity, we are only going to prove the case when $u_1 \perp u_2$ and $\Pr[s_1 = s_2 = 1] = \gamma$. The other case can be proved identically.

G.1 Stage 1

In this stage, by the induction hypothesis we know that $\|v_i^{(t)}\|_2 \leq \tilde{O}(1/\sqrt{d})$. Therefore, the update of $w_i^{(t)}$ can be approximate as:

Lemma G.1. *For every $t \leq T_{B,1}$, we know that: when the random samples are (X, z):*
$$w_i^{(t+1)} = w_i^{(t)} + \eta_D a^{(0)} \left(1 \pm \frac{1}{\mathrm{polylog}(d)}\right) \frac{3}{2} \langle w_i^{(t)}, X \rangle^2 X \pm \eta_D \tilde{O}\left(\frac{1}{d^{1.5}}\right) \tag{51}$$

Moreover, we have that if $z_i = 1$:
$$v_i^{(t+1)} = v_i^{(t)} + \eta_G a^{(0)} \frac{3}{2} \left(1 \pm \frac{1}{\mathrm{polylog}(d)}\right) \sum_{j \in [m_D]} \left(\langle w_j^{(0)}, G^{(t)}(z) \rangle \pm \frac{1}{C^{0.5} d}\right)^2 w_j^{(t)} \tag{52}$$
$$= v_i^{(t)} + \eta_G a^{(0)} \frac{3}{2} \left(1 \pm \frac{1}{\mathrm{polylog}(d)}\right) \sum_{j \in [m_D]} \left(\langle w_j^{(0)}, G^{(t)}(z) \rangle \pm \frac{1}{C^{0.5} d}\right)^2 w_j^{(0)} \pm \eta_G O\left(\frac{1}{C^{0.5} d^2}\right) \tag{53}$$

Taking expectation of the above Lemma, we can easily conclude that:
$$\mathbb{E}[w_i^{(t+1)}] = \mathbb{E}[w_i^{(t)}] + \eta_D a^{(0)} \left(1 \pm \frac{1}{\mathrm{polylog}(d)}\right) \frac{3}{4} \left(\langle w_i^{(t)}, u_1 \rangle^2 u_1 + \langle w_i^{(t)}, u_2 \rangle^2 u_2 + \Theta(\gamma) \langle w_i^{(t)}, u_1 + u_2 \rangle^2 (u_1 + u_2)\right) \tag{54}$$
$$\pm \eta_D \tilde{O}\left(\frac{1}{d^{1.5}}\right) \tag{55}$$

and
$$\mathbb{E}[\langle w_j^{(0)}, G^{(t)}(z) \rangle^2 \mid z_i = 1] = \langle w_j^{(0)}, v_i^{(t)} \rangle^2 \pm O\left(\frac{1}{m_G \,\mathrm{polylog}(d)} \sum_{i \in [m_G]} |\langle w_j^{(0)}, v_i^{(t)} \rangle|\right)^2 \tag{56}$$

Therefore, let $\zeta_t = \max_{i \in [m_G], j \in [m_D]} \langle v_i^{(t)}, w_j^{(0)} \rangle$, $\Upsilon_t = \max_{j \in [m_D], \ell \in [2]} \langle w_j^{(t)}, u_\ell \rangle$, we have that:
$$\mathbb{E}[\Upsilon_{t+1}] = \Upsilon + \eta_D a^{(0)} \frac{3}{4} \left(1 \pm \frac{1}{\mathrm{polylog}(d)}\right) \Upsilon_t^2 \tag{57}$$
$$\mathbb{E}[\zeta_{t+1}] = \zeta_t + \eta_G a^{(0)} \frac{3}{2m_G} \left(1 \pm \frac{1}{\mathrm{polylog}(d)}\right) \zeta_t^2 \tag{58}$$

Proof of Lemma G.1. By the gradient formula, we have:

$$\nabla_{w_i} L(X,z) = -\operatorname{Sigmoid}(-f(X)) a\sigma'(\langle w_i, X\rangle)X + \operatorname{Sigmoid}(f(G(z))) a\sigma'(\langle w_i, G(z)\rangle)G(z)$$

$$\nabla_{v_i} L(X,z) = -1_{z_i=1}\operatorname{Sigmoid}(f(G(z))) a \sum_{j\in[m_D]} \sigma'(\langle w_i, G(z)\rangle)w_i$$

At iteration t, by induction hypothesis, we have that $a^{(t)} = a^{(0)}(1 \pm 1/C)$.

Moreover, by the induction hypothesis again, we have hat $|f(X)| = \tilde{O}(d^{-1.5})$ and $\|G(z)\|_2 \leq O(d^{-0.5})$. Together with $\|w_i^{(t)}\|_2 = \tilde{O}(1)$, this implies that

$$\|\operatorname{Sigmoid}(f(G(z))) a\sigma'(\langle w_i, G(z)\rangle)G(z)\|_2 = \tilde{O}(d^{-1.5})$$

This proves the update formula for $w_i^{(t)}$. As for v_i, we observe that by the induction hypothesis and notice that w.h.p. over the randomness of initialization, $|\langle v_i^{(0)}, u_\ell\rangle| \leq \frac{\log d}{d}$, therefore, we can conclude that

$$\langle w_j^{(t)}, G^{(t)}(z)\rangle = \langle w_j^{(0)}, G^{(t)}(z)\rangle \pm \tilde{O}\left(\frac{1}{d^{1.35}}\right) \pm O\left(\frac{\log d}{Cd}\right) = \langle w_j^{(0)}, G^{(t)}(z)\rangle \pm \frac{1}{C^{0.5}d} \quad (59)$$

Note that by induction hypothesis, $\|w_j^{(t)} - w_j^{(0)}\|_2 \leq \frac{1}{C}$ and $\langle w_j^{(0)}, G^{(t)}(z)\rangle \leq \frac{m_G(C^{0.1}+\log d)}{d}$. This implies that

$$\langle w_j^{(t)}, G^{(t)}(z)\rangle^2 w_j^{(t)} = \left(\langle w_j^{(0)}, G^{(t)}(z)\rangle \pm \frac{1}{C^{0.5}d}\right)^2 w_j^{(t)} \quad (60)$$

$$= \left(\langle w_j^{(0)}, G^{(t)}(z)\rangle \pm \frac{1}{C^{0.5}d}\right)^2 w_j^{(0)} + O\left(\frac{1}{C^{0.5}d^2}\right) \quad (61)$$

□

Now, apply Lemma E.4 and the fact that w.p. $1-o(1)$, $\zeta_0 = \frac{\operatorname{polyloglog}(d)}{d}$, we have that:

Lemma G.2.

$$\sum_{t\leq T_1} \eta_G \zeta_t^2 \leq O\left(\frac{m_G \operatorname{polyloglog}(d)}{a^{(0)}d}\right) \quad (62)$$

In the end, we can show the following Lemma:

Lemma G.3. *When $t = T_{B,1}$, we have that: for both $\ell \in [2]$,*

$$\alpha(w_{i^*}, u_\ell, t) = \frac{1}{\operatorname{polylog}(d)}$$

Proof of Lemma G.3. By the update formula in Eq equation 54, and the fact that $\Pr[X = u_1 + u_2] \geq \frac{1}{\operatorname{polylog}(d)}$ and the induction hypothesis, we know that for $i = i^*$, for $t \leq T_{B,1}$ we have that:

$$\alpha(w_i, u_\ell, t+1) \geq \alpha(w_i, u_\ell, t) + \tilde{\Omega}(\eta_D) \times \left(\alpha(w_i, u_{3-\ell}, t) - \frac{1}{d}\right)^2$$

This implies that at the end of Stage 1, when $\alpha(w_i, u_{3-\ell}, t) \geq \frac{1}{C}$, we must have $\alpha(w_i, u_\ell, t) \geq \tilde{\Omega}(1)$ as well. □

G.2 Stage 2 and Stage 3

At this stage, by the induction hypothesis, we can approximate the function value as:

$$\sum_{i\in[m_D]} \sigma\left(\langle w_i^{(t)}, X\rangle\right) = \langle w_{i^*}^{(t)}, X\rangle^3 \pm \tilde{O}\left(\frac{1}{d^{1.5}}\right) \tag{63}$$

$$\left|\sum_{i\in[m_D]} \sigma\left(\langle w_i^{(t)}, G^{(t)}(z)\rangle\right)\right| \le \tilde{O}\left(\|G^{(t)}(z)\|_2\right)^3 \le \tilde{O}\left(\frac{1}{d^{1.45}}\right) \tag{64}$$

Therefore, at this stage, we can easily approximate the update of $W_D^{(t)}$ as:

Lemma G.4. *When the sample is (X, z), we have: for every $t \in (T_{B,1}, T_{B,3}]$, the following holds:*

$$a^{(t+1)} = a^{(t)} + \eta_D\left(1 \pm \tilde{O}\left(\frac{1}{d}\right)\right)\mathrm{Sigmoid}\left(-a^{(t)}\langle w_{i^*}^{(t)}, X\rangle^3 - b^{(t)}\right)\langle w_{i^*}^{(t)}, X\rangle^3 \tag{65}$$

$$\pm\, \eta_D\tilde{O}\left(\frac{1}{d^{1.45}}\right)\mathrm{Sigmoid}\left(b^{(t)}\right) \tag{66}$$

$$w_i^{(t+1)} = w_i^{(t)} + 3\eta_D\left(1 \pm \tilde{O}\left(\frac{1}{d}\right)\right)\mathrm{Sigmoid}\left(-a^{(t)}\langle w_i^{(t)}, X\rangle^3 - b^{(t)}\right)a^{(t)}\langle w_i^{(t)}, X\rangle^2 X \tag{67}$$

$$\pm\, \eta_D\tilde{O}\left(\frac{1}{d^{1.45}}\right)\mathrm{Sigmoid}\left(b^{(t)}\right) \tag{68}$$

$$b^{(t)} = b^{(t)} + \eta_D\tau_b\left(1 \pm \tilde{O}\left(\frac{1}{d}\right)\right)\mathrm{Sigmoid}\left(-a^{(t)}\langle w_{i^*}^{(t)}, X\rangle^3 - b^{(t)}\right) \tag{69}$$

$$-\, \eta_D\tau_b\left(1 \pm \tilde{O}\left(\frac{1}{d^{1.45}}\right)\right)\mathrm{Sigmoid}\left(b^{(t)}\right) \tag{70}$$

Moreover, the update formula also let us bound $a^{(t)}, \alpha(w_{i^*}, u_1, t)$ as:

Lemma G.5. *Let α_t, a_t be updated as: for $t = T_{B,2}$, $\alpha_t = \alpha(w_{i^*}, u_1, t)$ and $a_t = a^{(t)}$, such that*

$$a_{t+1} = a_t + \eta_G\,\mathrm{Sigmoid}(-a_t\alpha_t^3 - b_t)\alpha_t^3$$

$$\alpha_{t+1} = \alpha_t + \frac{3}{2}\eta_G a_t\,\mathrm{Sigmoid}(-a_t\alpha_t^3 - b_t)\alpha_t^2$$

Where b_t be updated as: for $t = T_{B,2}$, $b_t = b^{(t)}$ and update as:

$$b_{t+1} = b_t - \eta_D\tau_b\,\mathrm{Sigmoid}(b_t)$$

Then we have: for every $t \in [T_{B,2}, T_{B,3}]$

$$a_t = \left(1 \pm \frac{1}{\mathrm{polylog}(d)}\right)a^{(t)}, \quad \alpha_t = \left(1 \pm \frac{1}{\mathrm{polylog}(d)}\right)\alpha(w_{i^*}, u_1, t)$$

$$\mathrm{Sigmoid}(b^{(t)}) = \left(1 \pm \frac{1}{\mathrm{polylog}(d)}\right)\mathrm{Sigmoid}(b_t)$$

Moreover, when $t = T_{B,3}$, we have: $a_t \le 0.819\alpha_t$.

Proof of Lemma G.5. By the update formula in Lemma G.4 and the bound in induction hypothesis (B.3.3), we can simplify the update of $a^{(t)}, b^{(t)}$ and $\alpha(w_i, u_\ell, t)$ as: for $i = i^*$, when $X = u_\ell$:

$$a^{(t+1)} = a^{(t)} + \eta_D \left(1 \pm \frac{1}{\text{polylog}(d)}\right) \text{Sigmoid}\left(-a^{(t)}\alpha(w_i, u_\ell, t)^3 - b^{(t)}\right) \alpha(w_i, u_\ell, t)^3 \tag{71}$$

$$\alpha(w_i, u_\ell, t+1) = \alpha(w_i, u_\ell, t) \tag{72}$$

$$+ 3\eta_D \left(1 \pm \frac{1}{\text{polylog}(d)}\right) \text{Sigmoid}\left(-a^{(t)}\alpha(w_i, u_\ell, t)^3 - b^{(t)}\right) a^{(t)}\alpha(w_i, u_\ell, t)^2 \tag{73}$$

$$b^{(t)} = b^{(t)} - \eta_D \tau_b \left(1 \pm \frac{1}{\text{polylog}(d)}\right) \text{Sigmoid}\left(b^{(t)}\right) \tag{74}$$

By the last inequality, we know that when $\text{Sigmoid}(b^{(t)}) > \left(1 + \frac{1}{\text{polylog}(d)}\right) \text{Sigmoid}(b_t)$, then $b^{(t)}$ must be decreasing faster than b_t, otherwise if $\text{Sigmoid}(b^{(t)}) > \left(1 - \frac{1}{\text{polylog}(d)}\right) \text{Sigmoid}(b_t)$, then b_t must be decreasing faster than $b^{(t)}$, which proves the bound of $b^{(t)}$. Moreover, the update formula of b_t also gives us that for every $t \leq \text{poly}(d)$, we have: $|b_t| = O(\log d)$. This implies that for every Z:

$$\text{Sigmoid}(Z + b^{(t)}) = \left(1 \pm \frac{1}{\text{polylog}(d)}\right) \text{Sigmoid}(Z + b_t)$$

To obtain the bound of $a^{(t)}$ and $\alpha(w_i^*, u_1, t)$, notice that when $X = u_1 + u_2$, we have that:

$$\text{Sigmoid}\left(-a^{(t)}(\alpha(w_i, u_1, t) + \alpha(w_i, u_2, t))^3 - b^{(t)}\right) \leq \min_{\ell \in [2]} \text{Sigmoid}\left(-a^{(t)}\alpha(w_i, u_\ell, t)^3 - b^{(t)}\right)$$

Therefore, we can conclude:

$$\mathbb{E}[a^{(t+1)}] = a^{(t)} \tag{75}$$

$$+ \eta_D \frac{1}{2} \left(1 \pm \frac{1}{\text{polylog}(d)}\right) \left(\sum_{\ell \in [2]} \text{Sigmoid}\left(-a^{(t)}\alpha(w_i, u_\ell, t)^3 - b_t\right) \alpha(w_i, u_\ell, t)^3\right) \tag{76}$$

$$\mathbb{E}[\alpha(w_i, u_\ell, t+1)] = \alpha(w_i, u_\ell, t) \tag{77}$$

$$+ \frac{3}{2}\eta_D \left(1 \pm \frac{1}{\text{polylog}(d)}\right) \text{Sigmoid}\left(-a^{(t)}\alpha(w_i, u_\ell, t)^3 - b_t\right) a^{(t)}\alpha(w_i, u_\ell, t)^2 \tag{78}$$

Using Lemma G.3 we can conclude that

$$\alpha(w_{i^*}, u_\ell, T_{B,1}) = \frac{1}{\text{polylog}(d)}$$

and now apply Lemma E.7, we have: $a^{(t)} = \Theta(\alpha(w_{i^*}, u_1, t))$ and

$$\alpha(w_{i^*}, u_1, t) = [\alpha(w_{i^*}, u_1, t)] \left(1 \pm \frac{1}{\text{polyloglog}(d)}\right)$$

This implies that:

$$\mathbb{E}[a^{(t+1)}] = a^{(t)} + \eta_D \left(1 \pm \frac{1}{\text{polylog}(d)}\right) \left(\text{Sigmoid}\left(-a^{(t)}\alpha(w_i, u_1, t)^3 - b_t\right) \alpha(w_i, u_\ell, t)^3\right) \tag{79}$$

$$\mathbb{E}[\alpha(w_i, u_1, t+1)] = \alpha(w_i, u_1, t) \tag{80}$$

$$+ \frac{3}{2}\eta_D \left(1 \pm \frac{1}{\text{polylog}(d)}\right) \text{Sigmoid}\left(-a^{(t)}\alpha(w_i, u_1, t)^3 - b_t\right) a^{(t)}\alpha(w_i, u_1, t)^2 \tag{81}$$

Apply Lemma E.7 again, we know that when

$$a^{(t)}\alpha(w_i, u_1, t)^3 > \left(1 \pm \frac{1}{\mathrm{polylog}(d)}\right) a_t \alpha_t^3$$

We must have that $a^{(t)} \geq a_t$ and $\alpha(w_i, u_1, t) > \alpha_t$. Therefore, apply Lemma E.6 we know that in this case:

$$\mathrm{Sigmoid}\left(-a^{(t)}\alpha(w_i, u_1, t)^3 - b_t\right) \alpha(w_i, u_\ell, t)^3 \leq \mathrm{Sigmoid}(-a_t\alpha_t^3 - b_t)\alpha_t^3$$

and

$$\mathrm{Sigmoid}\left(-a^{(t)}\alpha(w_i, u_1, t)^3 - b_t\right) a^{(t)}\alpha(w_i, u_1, t)^2 \leq \mathrm{Sigmoid}(-a_t\alpha_t^3 - b_t)a_t\alpha_t^2$$

Combine this with the update rule we can directly complete the proof.

□

The Lemma G.5 immediately implies that the $\alpha(w_{i^*}, u_\ell, t)$ will be balanced after a while:

Lemma G.6. *We have that for every $t \in [T_{B,2}, T_{B,3}]$, the following holds:*

$$\alpha(w_{i^*}, u_1, t) = [\alpha(w_{i^*}, u_1, t)]\left(1 \pm \frac{1}{\mathrm{polyloglog}(d)}\right)$$

and

$$\alpha(w_{i^*}, u_1, t) \geq \log^{0.1}(d)$$

Using Lemma G.6, we also have the Lemma that approximate the update of $v_i^{(t)}$ as:

Lemma G.7. *Let us define $\alpha(t) := \max_{j\in[m_G],\ell\in[2]}\langle v_j^{(t)}, u_\ell\rangle$. For every $t \in [T_{B,2}, T_{B,3}]$, we have: for $j \neq i^*$:*

$$\langle w_j^{(t)}, G^{(t)}(z)\rangle^2 w_j^{(t)} = \langle w_j^{(t)}, G^{(t)}(z)\rangle^2 w_j^{(0)} \pm \frac{C^2}{d^{0.5}}\langle w_j^{(t)}, G^{(t)}(z)\rangle^2 \tag{82}$$

For $j = i^*$:

$$\langle w_j^{(t)}, G^{(t)}(z)\rangle^2 w_j^{(t)} = \langle w_j^{(t)}, G^{(t)}(z)\rangle^2 \left(w_j^{(0)} + \alpha(w_j, u_1, t)u_1 + \alpha(w_j, u_2, t)u_2\right) \pm \frac{C}{d^{0.9}}\langle w_j^{(t)}, G^{(t)}(z)\rangle^2 \tag{83}$$

Now, for $\langle w_j^{(t)}, G^{(t)}(z)\rangle$ we have: For $j \neq i^*$:

$$\mathbb{E}_z[\langle w_j^{(t)}, G^{(t)}(z)\rangle^2 \mid z_i = 1] = \left(1 \pm \frac{1}{\mathrm{polylog}(d)}\right)\left(\langle w_j^{(0)}, v_i^{(t)}\rangle \pm \frac{C^2}{\sqrt{d}}\alpha(t)\right)^2 \pm \frac{\alpha(t)^2}{\mathrm{polylog}(d)} \tag{84}$$

For $j = i^*$:

$$\mathbb{E}_z[\langle w_j^{(t)}, G^{(t)}(z)\rangle^2 \mid z_i = 1] = \left(1 \pm \frac{1}{\mathrm{polyloglog}(d)}\right)\alpha(w_{i^*}, u_1, t)^2\left\langle (u_1 + u_2), v_i^{(t)}\right\rangle^2 \tag{85}$$

$$\pm \frac{\alpha(t)^2}{\mathrm{polyloglog}(d)}\alpha(w_{i^*}, u_1, t)^2 \tag{86}$$

Moreover, the update of Sigmoid can be approximate as:

$$\mathrm{Sigmoid}\left(f^{(t)}\left(G^{(t)}(z)\right)\right) = \left(1 \pm \tilde{O}\left(\frac{1}{d^{1.45}}\right)\right)\mathrm{Sigmoid}\left(b^{(t)}\right)$$

Proof of Lemma G.7. The first half of the lemma regarding $w_j^{(t)}$ follows trivially from the induction hypothesis, we only need to look at $\langle w_j^{(t)}, G^{(t)}(z)\rangle$.

We know that for $j \neq i^*$, we have that by the induction hypothesis,

$$\langle w_j^{(t)}, G^{(t)}(z)\rangle = \langle w_j^{(0)}, G^{(t)}(z)\rangle \pm \tilde{O}\left(\frac{1}{d^{1.35}}\right) \pm O\left(\frac{Cm_G\alpha(t)}{\sqrt{d}}\right) \tag{87}$$

For $j = i^*$, we have that

$$\langle w_j^{(t)}, G^{(t)}(z)\rangle = \langle w_j^{(0)}, G^{(t)}(z)\rangle + \alpha(w_j, u_1, t)\langle u_1, G^{(t)}(z)\rangle + \alpha(w_j, u_2, t)\langle u_2, G^{(t)}(z)\rangle \pm \tilde{O}\left(\frac{1}{d^{1.35}}\right) \tag{88}$$

$$= \alpha(w_j, u_1, t)\langle u_1 + u_2, G^{(t)}(z)\rangle \pm \frac{1}{\mathrm{polyloglog}(d)}\alpha(w_j, u_1, t)\alpha(t)\|z\|_1 \pm \tilde{O}\left(\frac{1}{d^{1.35}}\right) \tag{89}$$

$$= \alpha(w_j, u_1, t)\langle u_1 + u_2, v_i^{(t)}\rangle \pm \tilde{O}\left(\frac{1}{d^{1.35}}\right) \pm O(\alpha(w_j, u_1, t)\alpha(t))(\|z\|_1 - 1) \tag{90}$$

Taking expectation we can complete the proof.

□

With Eq equation 82 and Eq equation 83 in lemma G.7, together with the induction hypothesis, we immediately obtain

Lemma G.8. *For every $t \in [T_{B,2}, T_{B,3}]$, we have that: for every $i \in [m_G]$:*

$$v_i^{(t)} = v_i^{(T_{B,2})} + \sum_{\ell\in[2]} \alpha_{i,\ell}^{(t)} u_\ell + \sum_{j\in[m_D]} \beta_{i,j}^{(t)} w_j^{(0)} \pm \xi_{i,t} \tag{91}$$

Where $\alpha_{i,\ell}^{(t)}, \beta_{i,j}^{(t)} > 0$ and $\alpha_{i,\ell}^{(t)} = (1 \pm o(1))\alpha_{i,3-\ell}^{(t)}$; $\|\xi_{i,t}\|_2^2 \le \tilde{O}(1/d)\left(\sum_{\ell,j}(\alpha_{i,\ell}^{(t)})^2 + (\beta_{i,j}^{(t)})^2\right)$

We now can immediately control the update of $v_i^{(t)}$ using the following sequence:

Lemma G.9. *Let v_t be defined as: for $t = T_{B,2}$*

$$v_t = \max_{i\in[m_G]} \langle v_i^{(t)}, u_1 + u_2\rangle \left(1 + \frac{1}{\mathrm{polyloglog}(d)}\right)$$

and the update of v_t is given as: for α_t defined as in Lemma G.5

$$v_{t+1} = v_t + \frac{3}{m_G}\,\mathrm{Sigmoid}(b_t)\alpha_t^2 v_t^2$$

Then we must have: for every $t \in [T_{B,2}, T_{B,3}]$:

$$\max_{i\in[m_G]} \langle v_i^{(t)}, u_1 + u_2\rangle \le v_t \tag{92}$$

On the other hand, if for $t = T_{B,2}$,

$$v_t = \max_{i\in[m_G]} \langle v_i^{(t)}, u_1 + u_2\rangle \left(1 - \frac{1}{\mathrm{polyloglog}(d)}\right)$$

Then we must have: for every $t \in [T_{B,2}, T_{B,3}]$:

$$\max_{i\in[m_G]} \langle v_i^{(t)}, u_1 + u_2\rangle \ge v_t \tag{93}$$

Proof of Lemma G.9. In the setting of Lemma G.7, let us define $beta(t) := \max_{j \in [m_G]} \langle v_j^{(t)}, u_1 + u_2 \rangle$. We have that:

$$\beta(t+1) = \beta(t) + \eta_G \alpha_t \left(1 \pm \frac{1}{\operatorname{polylog}(d)}\right) \frac{3}{m_G} \beta(t)^2 \pm \frac{1}{\operatorname{polyloglog}(d)} \alpha(t)^2 \tag{94}$$

By the induction hypothesis we know that for all $j \in [m_G], \ell \in [2]$:

$$\langle v_j^{(t)}, u_\ell \rangle \geq \langle v_j^{(0)}, u_\ell \rangle - O\left(\frac{1}{d}\right) \geq -\frac{\log\log^2 d}{d} \tag{95}$$

This implies that $\beta(t) \geq \alpha(t) - \frac{\log\log^2(d)}{\sqrt{d}}$ and $\beta(T_{B,2}) \geq \frac{1}{d}$, $\alpha(T_{B,2}) \leq \frac{\operatorname{polyloglog}(d)}{\sqrt{d}}$. This implies that:

$$\beta(t+1) = \beta(t) + \eta_G \alpha_t \left(1 \pm \frac{1}{\operatorname{polyloglog}(d)}\right) \frac{3}{m_G} \beta(t)^2 \tag{96}$$

This completes the proof by applying Lemma E.1. □

Now, by the comparison Lemma E.1, we know that one of the following event would happen (depending on the initial value of v_t at iteration $T_{B,2}$):

Lemma G.10. *With probability $1 - o(1)$, one of the following would happen:*

1. $T_{B,3} \geq T_0$.
2. $T_{B,3} < T_0$, *moreover, at iteration $T_{B,3}$, we have that* $\operatorname{Sigmoid}(b_t) \geq \frac{1}{\operatorname{polylog}(d)}$.

In the end, we can easily derive an upper bound on the sum of Sigmoid as below, which will be used to prove the induction hypothesis.

Lemma G.11. *For every $t \in (T_{B,1}, T_{B,3}]$, we have that: for every X, z:*

$$\sum_{t \in (T_{B,1}, T_{B,3}]} \eta_D \operatorname{Sigmoid}\left(a^{(t)} \langle w_{i^*}^{(t)}, X \rangle^3 + b^{(t)}\right) = \tilde{O}(1) \tag{97}$$

$$\sum_{t \in (T_{B,1}, T_{B,3}]} \eta_D \tau_b \left(b^{(t)}\right) \leq \tilde{O}(1) \tag{98}$$

We will also show the following Lemma regarding all the $v_i^{(t)}$ at iteration $T_{B,3}$:

Lemma G.12. *For all $i \in [m_G]$, if we are in case 2 in Lemma G.10, we have that:*

$$\langle v_i^{(t)}, u_1 \rangle, \langle v_i^{(t)}, u_2 \rangle = \tilde{\Omega}(d^{-0.49}) \tag{99}$$

Proof of Lemma G.12. Since with probability at least $1/\operatorname{poly}(d)$, $z_i = z_j = 1$, so we have: By the update Lemma G.7 of v, we know that for all $j \in [m_G]$: Let $\alpha(t)$ be defined as in Lemma G.7:

$$\mathbb{E}[\alpha(v_j, u_\ell, t+1)] \geq \alpha(v_j, u_\ell, t) + \eta_G \frac{1}{\operatorname{polylog}(d)} \left(\alpha(t) - \tilde{O}\left(\frac{1}{d}\right)\right)^2 \pm \tilde{O}\left(\frac{1}{d^{0.8}}\right) \alpha(t)^2 \tag{100}$$

$$\mathbb{E}[\alpha(v_j, u_\ell, t+1)] \leq \alpha(v_j, u_\ell, t) + \eta_G \operatorname{polylog}(d) \alpha(t)^2 \tag{101}$$

By Lemma G.10 we know that $\alpha(t) = \tilde{\Theta}(d^{-0.49})$ at iteration $t = T_{B,3}$, which implies what we want to prove. □

G.3 Stage 4 and 5

In Stage 4 we can easily calculate that by induction hypothesis, for every $i \in [m_G]$ and for every $j \in [m_D]$, $j \neq i^*$:

$$|\langle v_i^{(t)}, w_j^{(t)} \rangle| \leq \tilde{O}\left(\frac{\|v_i^{(t)}\|_2}{\sqrt{d}} \right)$$

Let

$$S_{i,t} = \mathbb{E}_z \left[\text{Sigmoid}\left(a^{(t)} \sigma\left(\langle w_{i^*}^{(t)}, G^{(t)}(z) \rangle \right) + b^{(t)} \right) \mid z_i = 1 \right]$$

Note that by induction hypothesis, $|a^{(t)}|, b^{(t)} = O(\log(d))$. Which implies that as long as $\max_{i \in [m_G]} \|v_i^{(t)}\|_2 \leq \frac{1}{\text{polylog}(d)}$ or for all $i \in [m_G]$, $a^{(t)} \sigma'(\langle v_i^{(t)}, w_{i^*} \rangle) \geq \tilde{\Omega}(\log(d))$, we have that: for all z, z' we have that:

$$\text{Sigmoid}\left(a^{(t)} \sigma\left(\langle w_{i^*}^{(t)}, G^{(t)}(z) \rangle \right) + b^{(t)} \right) = \Theta(1) \times \text{Sigmoid}\left(a^{(t)} \sigma\left(\langle w_{i^*}^{(t)}, G^{(t)}(z') \rangle \right) + b^{(t)} \right)$$

We can immediately obtain the following Lemma:

Lemma G.13. *The update of $v_i^{(t)}$ is given as:*

$$\mathbb{E}[v_i^{(t+1)}] = v_i^{(t)} + \tilde{\Theta}(\eta_G) a^{(t)} S_{i,t} \left(\left(\sigma'(\langle w_{i^*}^{(t)}, v_i^{(t)} \rangle + \sum_{j \in [m_G]} \gamma_{i,j,t} \sigma'(\langle w_{i^*}^{(t)}, v_j^{(t)} \rangle) \right) w_i^{(t)} \pm \tilde{O}\left(\frac{\|v_i^{(t)}\|_2}{\sqrt{d}} \right)^2 \right) \tag{102}$$

Where $\gamma_{i,j,t} > 0$; $\gamma_{i,j,t} = \tilde{\Theta}(1)$ if $\max_{i \in [m_G]} \|v_i^{(t)}\|_2 \leq \frac{1}{\text{polylog}(d)}$ or for all $i \in [m_G]$, $a^{(t)} \sigma'(\langle v_i^{(t)}, w_{i^*} \rangle) \geq \tilde{\Omega}(\log(d))$, and $\gamma_{i,j,t} = \tilde{O}(1)$ otherwise.

Here the additional $\sigma'(\langle w_{i^*}^{(t)}, v_j^{(t)} \rangle)$ part comes from $\Pr[z_i, z_j = 1] = \frac{1}{\text{polylog}(d)}$. The remaining part of this stage follows from simply apply Lemma E.4.

In stage 5, we bound the update of $a^{(t)}, b^{(t)}$ as:

Let

$$S_t = \mathbb{E}_z \left[\text{Sigmoid}\left(a^{(t)} \sigma\left(\langle w_{i^*}^{(t)}, G^{(t)}(z) \rangle \right) + b^{(t)} \right) \right]$$

In this stage, with the induction hypothesis, we can easily approximate the sigmoid as:

Lemma G.14. *For $t \geq T_{B,4}$, the sigmoid can be approximate as: For every X, z*

$$\text{Sigmoid}(-f^{(t)}(X)) = \left(1 \pm \frac{1}{\text{polylog}(d)} \right) \text{Sigmoid}(-b^{(t)})$$

$$\text{Sigmoid}(f^{(t)}(G^{(t)}(z))) = \left(1 \pm \frac{1}{\text{polylog}(d)} \right) \text{Sigmoid}\left(a^{(t)} \sigma\left(\langle w_{i^*}^{(t)}, G^{(t)}(z) \rangle \right) + b^{(t)} \right)$$

Then by the update rule, we can easily conclude that:

Lemma G.15. *For $t \geq T_{B,4}$, the update of $a^{(t)}, b^{(t)}$ is given as:*

$$a^{(t+1)} = a^{(t)} + \tilde{O}(\eta_D) \, \text{Sigmoid}\left(-b^{(t)} \right) - \tilde{\Omega}(\eta_D) S_t \Lambda^2 d^{1/4}$$

$$\mathbb{E}[b^{(t+1)}] = b^{(t)} + \eta_D \tau_b \left(1 \pm \frac{1}{\text{polylog}(d)} \right) \text{Sigmoid}\left(-b^{(t)} \right) - \eta_D \tau_b \left(1 \pm \frac{1}{\text{polylog}(d)} \right) S_t$$

This Lemma, together with the induction hypothesis, implies that:

Lemma G.16. *We have:*

$$\sum_{t \geq T_{B,4}} S_t \leq \tilde{O}\left(\frac{1}{\eta_D \tau_b \Lambda^2 d^{1/4}}\right) \tag{103}$$

$$\sum_{t \geq T_{B,4}} \text{Sigmoid}(b^{(t)}) \leq \tilde{O}\left(\frac{1}{\eta_D \tau_b}\right) \tag{104}$$

Proof of Lemma G.16. Let us denote $R = \sum_{t=T_{B,4}}^{T_0} \text{Sigmoid}\left(-b^{(t)}\right)$ and $Q = \sum_{t=T_{B,4}}^{T_0} S_t$

Sum the update up for $t = T_{B,4}$ to T_0, we have that:

$$a^{(T_0)} - a^{(T_{B,4})} = \tilde{O}(\eta_D)R - \tilde{\Omega}(\eta_D)Q\Lambda^2 d^{1/4} \tag{105}$$

$$\mathbb{E}[b^{(T_0)}] - b^{(T_{B,4})} = \Theta(\eta_D \tau_b)R - \Theta(\eta_D \tau_b)Q \tag{106}$$

By the induction hypothesis that $|a^{(t)}| \leq \frac{\tilde{O}(1)}{\Lambda^2 d^{1/4}}$ and $|b^t| = \tilde{O}(1)$, we have that:

$$|\tilde{O}(\eta_D)R - \tilde{\Omega}(\eta_D)Q\Lambda^2 d^{1/4}| \leq \frac{\tilde{O}(1)}{\Lambda^2 d^{1/4}} \tag{107}$$

$$|\Theta(\eta_D \tau_b)R - \Theta(\eta_D \tau_b)Q| \leq \tilde{O}(1) \tag{108}$$

Thus, we have:

$$\tilde{\Omega}(\eta_D)Q\Lambda^2 d^{1/4} \leq \tilde{O}(\eta_D)R + \frac{\tilde{O}(1)}{\Lambda^2 d^{1/4}} \leq \tilde{O}(\eta_D)\left(\frac{1}{\eta_D \tau_b} + Q\right) + \frac{\tilde{O}(1)}{\Lambda^2 d^{1/4}} \tag{109}$$

Therefore we have that $\tilde{\Omega}(\eta_D)Q\Lambda^2 d^{1/4} \leq \tilde{O}\left(\frac{1}{\eta_D \tau_b} + \frac{1}{\eta_D \Lambda^2 d^{1/4}}\right)$, which implies that

$$Q \leq \tilde{O}\left(\frac{1}{\eta_D \tau_b \Lambda^2 d^{1/4}}\right)$$

Similarly, we can show that

$$R \leq \tilde{O}\left(\frac{1}{\eta_D \tau_b} + Q\right) \leq \tilde{O}\left(\frac{1}{\eta_D \tau_b} + \frac{1}{(\Lambda^2 d^{1/4})^2} + \frac{1}{\Lambda^2 d^{1/4}} R\right) \tag{110}$$

This implies that $R \leq \tilde{O}\left(\frac{1}{\eta_D \tau_b}\right)$.

□

G.4 Proof of the induction hypothesis and the final theorem

The final theorem follows immediately from the induction hypothesis (v part) together with Lemma G.8.

Now it remains to prove the induction hypothesis. We will assume that all the hypothesises are true until iteration t, then we will prove that they are true at iteration $t + 1$.

Stage 1 .

To prove the induction hypothesis at Stage 1, for w, we have that by Lemma G.1, we know that: for $* \neq u_1, u_2$,

$$|\alpha(w_i, *, t+1) - \alpha(w_i, *, t)| \leq \eta_D \tilde{O}\left(\frac{1}{d^{1.5}}\right) \tag{111}$$

By $T_1 \leq O\left(\frac{\sqrt{d}}{\eta_D a^{(0)}}\right)$ we can conclude that

$$|\alpha(w_i, *, t+1) - \alpha(w_i, *, 0)| \leq \eta_D \tilde{O}\left(\frac{1}{d^{1.5}}\right) \times O\left(\frac{\sqrt{d}}{\eta_D a^{(0)}}\right) \leq \frac{1}{d^{0.9}} \tag{112}$$

On the v part, again by Lemma G.1, we know that for $* \notin \{w_j\}_{j\in[m_D]}$:

$$|\alpha(v_i, *, t) - \alpha(v_i, *, 0)| \leq \eta_G \left(\frac{1}{C^{0.5}d^2}\right) \times O\left(\frac{\sqrt{d}}{\eta_D a^{(0)}}\right) \leq \frac{1}{d} \tag{113}$$

On the other hand, we know that for w_j:

$$|\alpha(v_i, w_j, t+1) - \alpha(v_i, w_j, t)| \leq \eta_G \left(1 + \frac{1}{\mathrm{polylog}(d)}\right) \frac{3}{2m_G} \zeta_t^2 \tag{114}$$

Apply Lemma G.2 we complete the proof using Lemma E.3.

As for the $a^{(t)}, b^{(t)}$ part, we know that:

$$|a^{(t+1)} - a^{(t)}| \leq O\left(\eta_D m_D \Upsilon_t^3\right), |b^{(t)}| \leq \tau_b \eta_D T_1 \tag{115}$$

Combine with the update rule of Υ in Eq equation 57, we complete the proof.

Stage 2 and 3 For the w part, we know that by Lemma G.4, we have that for every $* \neq u_1, u_2$

$$|\alpha(w_i, *, t+1) - \alpha(w_i, *, t)| \leq \eta_D \tilde{O}\left(\frac{1}{d^{1.45}}\right) \mathrm{Sigmoid}(b^{(t)}) \tag{116}$$

Now, by Lemma G.11 we have that:

$$\sum_{t\in(T_{B,1}, T_{B,3}]} \eta_D \tau_b \left(b^{(t)}\right) \leq \tilde{O}(1)$$

This implies that

$$|\alpha(w_i, *, t+1) - \alpha(w_i, *, T_{B,1})| \leq \eta_D \tilde{O}\left(\frac{1}{d^{1.45}}\right) \times \frac{1}{\eta_b \eta_D} \leq \frac{1}{d^{0.9}}$$

For the v part for $t \leq T_{B,2}$, since $T_{B,2} - T_{B,1} = \tilde{O}(d^{o(1)}/\eta_D)$, we can easily prove it for $t \leq T_{B,2}$ as in stage 1. On the other hand, for $t \in (T_{B,2}, T_{B,3}]$: By Lemma G.7 and Lemma G.6, we have that define

$$\alpha(t) := \max_{j\in[m_G],\ell\in[2]} \langle v_j^{(t)}, u_\ell \rangle, \quad \beta(t) := \max_{j\in[m_G],j'\in[m_G],j\neq j';i\in[m_D],i\neq i^*} |\langle v_j^{(t)}, w_i^{(0)} \rangle| + |\langle v_j^{(t)}, v_{j'}^{(0)} \rangle|$$

We have that:

$$\mathbb{E}[\alpha(t+1)] \geq \alpha(t) + \eta_G \Omega\left(\frac{1}{m_G}\right) \mathrm{Sigmoid}(b^{(t)}) \alpha(t)^2 \log^{0.1}(d) \tag{117}$$

and

$$\mathbb{E}[\beta(t+1)] \leq \beta(t) + \eta_G O\left(\frac{1}{m_G}\right) \mathrm{Sigmoid}(b^{(t)}) \left(\beta(t)^2 + \frac{C^2}{\sqrt{d}} \alpha(t)^2\right) \tag{118}$$

By Lemma D.2 and Lemma E.1 we can show that $\beta(t) = O\left(\beta(0) + \frac{C^2}{\sqrt{d}}\alpha(t)\right)$, which complete the proof that for all $* \neq w_{i^*}, u_1, u_2$:

$$|\alpha(v_j, *, t)| \leq \frac{C^3}{\sqrt{d}} \|v_j^{(t)}\|_2$$

.

Stage 4 and 5 At stage 4 we simply use Lemma E.4, the only remaining part is to show that $|b^{(t)}| = O(\log(d))$. To see this, we know that by the update formula:

$$\nabla_b L(X, z) = -\operatorname{Sigmoid}(-f(X)) + \operatorname{Sigmoid}(f(G(z)))$$

By our induction hypothesis, we know that $a^{(t)}\left(\sum_{i\in[m_D]}\sigma(\langle w_i^{(t)}, X\rangle)\right) > 0$

and $a^{(t)}\left(\sum_{i\in[m_D]}\sigma(\langle w_i^{(t)}, G(z)\rangle)\right) > 0$. Therefore, $b < O(\log(d))$ is immediate. Now it remains to show that $b > -O(\log d)$: By the update formula, we have:

$$-b^{(t+1)} \le -b^{(t)} + \eta_D\tau_b \sum_{i\in[m_G]} S_{i,t}$$

and by Lemma G.13 and the proof in Lemma E.4, we have that:

$$\sum_{i\in[m_G]}\langle v_i^{(t+1)}, w_{i^*}^{(0)}\rangle \ge \sum_{i\in[m_G]}\langle v_i^{(t)}, w_{i^*}^{(0)}\rangle + a^{(t)}\tilde{\Omega}(\eta_G)\left(\sum_{i\in[m_G]} S_{i,t}\right)\left(\sum_{i\in[m_G]}\sigma'(\langle v_i^{(t)}, w_{i^*}^{(0)}\rangle)\right) \tag{119}$$

Compare this two updates we can easily obtain that $|b^{(t)}| = O(\log(d))$.

At stage 5, we have that since $|a^{(t)}| = \tilde{O}\left(\frac{1}{d^{1/4}\Lambda^2}\right)$: For every $j \in [m_G], i \in [m_D]$

$$\|v_j^{(t+1)} - v_j^{(t)}\|_2 \le \tilde{O}(\eta_G) S_t \Lambda^2 \times \frac{1}{\Lambda^2 d^{1/4}} \tag{120}$$

$$\|w_i^{(t+1)} - w_i^{(t)}\|_2 \le \tilde{O}(\eta_D)\operatorname{Sigmoid}(-b^{(t)})\frac{1}{d^{1/4}\Lambda^2} \tag{121}$$

Apply Lemma G.16 we have that:

$$\|v_j^{(t+1)} - v_j^{(T_{B,4})}\|_2 \le \tilde{O}(\eta_G) \times \frac{1}{d^{1/4}} \times \frac{1}{\eta_D\tau_b\Lambda^2 d^{1/4}} \le d^{0.15} \tag{122}$$

$$\|w_i^{(t+1)} - w_i^{(T_{B,4})}\|_2 \le \tilde{O}(\eta_D)\frac{1}{d^{1/4}\Lambda^2} \times \frac{1}{\eta_D\tau_b} \le \frac{1}{d^{0.1}} \tag{123}$$

Which proves the induction hypothesis.

H Proof of the learning process in other cases

We now consider other cases, in case 1 of Lemma D.1, the proof is identical to case 2, the only difference is at Stage 3, we have that $T_{B,3} > T_0$.

In case 2, the Stage 1 is identical to the Stage 1, 2, 3 in the balanced case. For Stage 3, its identical to Stage 4 in the balanced case (the only difference is to apply Lemma E.5 and the case 2 of Lemma E.2 instead of Lemma E.4). For Stage 4, its identical to Stage 5 in the balanced case.

At Stage 2, by the induction hypothesis, we know that for $j \neq i_G$, we have that $|\langle v_j^{(t)}, w_j^{(t)}\rangle| \le \tilde{O}\left(\frac{1}{C}\|v_j^{(t)}\|_2\right)$. Thus, we can approximate the update of w, v as:

$$w_i^{(t+1)} = w_i^{(t)} \pm \tilde{O}(\eta_D \sum_{j\in[m_G]}\|v_j^{(t)}\|_2^2)\Lambda \tag{124}$$

$$v_j^{(t+1)} = v_j^{(t)} + \tilde{\Theta}(\eta_G \sum_{j\in[m_G]}\|v_j^{(t)}\|_2^2) w_{i_G}^{(0)} \pm \frac{1}{C^2}\tilde{\Theta}(\eta_G \sum_{j\in[m_G]}\|v_j^{(t)}\|_2^2) \tag{125}$$

Using the fact that $\eta_G \ge \tilde{\Omega}(\sqrt{d})\eta_D$ in case 3 we immediately proves the induction hypothesis.

The proof of the theorem follows immediately from the induction hypothesis on v in this case v only learns noises (linear combinations of $w_i^{(0)}$).

I NORMALIZED SGD

In this section we look at the update of normalized SGD.

Let us define:

$$i_1^* = \arg\max_{i\in[m_D]}\{\langle w_i^{(0)}, u_1\rangle\}$$

$$i_2^* = \arg\max_{i\in[m_D]}\{\langle w_i^{(0)}, u_2\rangle\}$$

Let us define:

$$g_j^* = \arg\max_{i\in[m_D]}\{\langle v_j^{(0)}, w_i^{(0)}\rangle\}$$

Then we first show the following Lemma about initialization:

Lemma I.1. *With probability at least $1 - o(1)$ over the randomness of the initialization, the following holds:*

1. *For all $\ell \in [2]$, for all $i \in [m_D]$ such that $i \neq i_\ell^*$, we have:*

$$\langle w_{i_\ell^*}^{(0)}, u_\ell\rangle \geq \left(1 - \frac{1}{\mathrm{polyloglog}(d)}\right)\langle w_i^{(0)}, u_\ell\rangle$$

2. *For all $j \in [m_G]$, we have that for all $i \in [m_D]$ such that $i \neq g_j^*$,*

$$\langle v_j^{(0)}, w_{g_j^*}^{(0)}\rangle \geq \left(1 - \frac{1}{\log^4(d)}\right)\langle v_j^{(0)}, w_i^{(0)}\rangle$$

3. $\{g_j^*\}_{j\in[m_G]} = [m_D]$.

We now divide the training stage into two: For a sufficiently large $C = \mathrm{polylog}(d)$, consider the case when $\eta_G = \eta_D * C^{-0.6}$.

1. Stage 1: When both $\alpha(w_{i*_1}, u_1, t), \alpha(w_{i*_2}, u_2, t) \leq \frac{1}{C^{0.95}}$. Call this iteration $T_{N,1}$.
2. Stage 2: After $T_{N,1}$, before T_1

I.1 INDUCTION HYPOTHESIS

We will use the following induction hypothesis: for a

Stage 1: for every $t \leq T_{N,1}$: Let $\alpha(t) := \max_{\ell\in[2]} \alpha(w_{i_\ell^*}, u_\ell, t)$, $\beta(t) := \max_{i\in[m_G]} \alpha(v_i, w_{g_i^*}, t)$.

1. Domination: For every $i \in [m_G]$, we have:

$$|\alpha(v_i, *, t) - \alpha(v_i, *, 0)| \leq \min\left\{\frac{1}{C}\alpha(t), \beta(t)\right\}$$

For every $i \in [m_D]$, $i \neq i_1^*, i_2^*$, we have that for $* \neq u_1, u_2$:

$$|\alpha(w_i, *, t) - \alpha(w_i, *, 0)| \leq \frac{1}{C}\alpha(t)$$

and

$$|\alpha(w_i, u_1, t)|, |\alpha(w_i, u_2, t)| \leq \alpha(t)$$

For i_1^*, we have that for every $* \neq u_1$,

$$|\alpha(w_{i_1^*}, *, t) - \alpha(w_{i_1^*}, *, 0)| \leq \frac{1}{C}\alpha(t)$$

For i_2^*, we have that for every $* \neq u_2$,

$$|\alpha(w_{i_2^*}, *, t) - \alpha(w_{i_2^*}, *, 0)| \leq \frac{1}{C}\alpha(t)$$

2. (N.1.2): Growth rate: we have that for every $i \in [m_D]$
$$\alpha(w_{i_\ell^*}, u_\ell, t) \in \left(\Omega\left(\frac{1}{m_D}\right), 1\right)\eta_D t$$
and for every $i \in [m_G]$:
$$\alpha(v_i, w_{g_i^*}, t) \in \left(\Omega\left(\frac{1}{m_G^2}\right), 1\right)\eta_G t$$
Therefore by our choice of η_D, η_G we have that $\beta(t) \in C^{-0.6}\left[\frac{1}{\log^5 d}, \log^5 d\right] \times \alpha(t)$.
3. $a^{(t)} = [0.5, 1]a^{(0)}, |b^{(t)}| \leq \frac{1}{d^{0.1}}$.

Stage 2: We maintain: For every $t \in [T_{N,1}, T_1]$:

1. (N.1.2) still holds.
2. $\alpha(w_{i_\ell^*}, u_\ell, t) \in \left[\frac{1}{C}, \text{polyloglogloglog}(d)\right]$, $\beta(t) \in C^{-0.6}\left[\frac{1}{\log^5 d}, \log^5 d\right] \times \alpha(t)$, $a^{(t)} = \Omega(\alpha(w_{i_\ell^*}, u_\ell, t))$.
3. w_i's are good: For every $i \neq i_1^*, i_2^*$, for every $*$:
$$|\alpha(w_i, *, t) - \alpha(w_i, *, 0)| \leq \frac{1}{C}\alpha(t)$$
and for $\ell \in [2]$: for every $* \neq u_\ell$, we have:
$$|\alpha(w_i, *, t) - \alpha(w_i, *, 0)| \leq \frac{1}{C}\alpha(t)$$
4. v_i's are good: For every $i \in [m_G]$ and every $j \in [m_D]$, $j \neq g_i^*$:
$$\langle v_i^{(t)}, w_{g_i^*}^{(t)}\rangle \geq C^{0.9}|\langle v_i^{(t)}, w_j^{(t)}\rangle|$$
and for $g_i^* \neq i_\ell^*$, we have that:
$$\langle v_i^{(t)}, w_{g_i^*}^{(t)}\rangle \geq C^{0.9}|\langle v_i^{(t)}, u_\ell\rangle|$$
For $g_i^* = i_\ell$, we have that $\langle v_i^{(t)}, u_\ell\rangle \geq -\frac{1}{C^{0.5}}\beta(t)$

I.2 Stage 1 training

With the induction hypothesis, we can show the following Lemma:

Lemma I.2. *For $t \leq T_{N,1}$, for $\varepsilon_t := \frac{(\alpha(t)+d^{-0.5})^2}{C^{1.5}} + C^{0.5}(\alpha(t) + d^{-0.5})^3$, when the sample is $X \in \{u_1, u_2\}$, the update of $w_i^{(t)}$ can be approximate as:*
$$w_i^{(t+1)} = w_i^{(t)} + \eta_D\left(1 \pm \frac{1}{\text{polylog}(d)}\right)\frac{\sigma'(\langle w_i^{(t)}, X\rangle)X \pm \varepsilon_t}{\sqrt{\sum_{j\in[m_D]}\sigma'(\langle w_j^{(t)}, X\rangle)^2\|X\|_2^2}} \tag{126}$$

Which can be further simplified as:
$$\mathbb{E}[\langle w_i^{(t+1)}, u_\ell\rangle] = \langle w_i^{(t)}, u_\ell\rangle + \eta_D\frac{1}{2}\left(1 \pm \frac{1}{\text{polylog}(d)}\right)\frac{\sigma'(\langle w_i^{(t)}, u_\ell\rangle) \pm \varepsilon_t}{\sqrt{\sum_{j\in[m_D]}\sigma'(\langle w_j^{(t)}, u_\ell\rangle)^2}} \pm \eta_D\gamma \tag{127}$$

When $z = e_i$, the update of v can be approximate as: For $\delta_t := O\left(\frac{1}{C^{0.94}}(\beta(t) + \frac{1}{d})^2\right)$:
$$v_i^{(t+1)} = v_i^{(t)} + \eta_G\left(1 \pm \frac{1}{\text{polylog}(d)}\right)\frac{\sum_{j\in[m_D]}\sigma'(\langle v_i^{(t)}, w_j^{(0)}\rangle)w_j^{(0)} \pm \delta_t}{\|\sum_{j\in[m_D]}\sigma'(\langle v_i^{(t)}, w_j^{(0)}\rangle)w_j^{(0)}\|_2} \tag{128}$$

Where we have:
$$\langle v_i^{(t)}, w_j^{(t)}\rangle = \langle v_i^{(t)}, w_j^{(0)}\rangle \pm \frac{1}{C^{0.9}}\beta(t)$$

Proof of the update Lemma I.2. By the induction hypothesis, We have that:

$$\langle v_i^{(t)}, w_j^{(t)} \rangle = \langle v_i^{(t)}, w_j^{(0)} \rangle + \langle v_i^{(t)}, w_j^{(t)} - w_j^{(0)} \rangle \tag{129}$$

$$= \langle v_i^{(t)}, w_j^{(0)} \rangle + \langle v_i^{(0)}, w_j^{(t)} - w_j^{(0)} \rangle + \langle v_i^{(t)} - v_i^{(0)}, w_j^{(t)} - w_j^{(0)} \rangle \tag{130}$$

$$= \langle v_i^{(t)}, w_j^{(0)} \rangle \pm \tilde{O}\left(\frac{1}{\sqrt{d}}\alpha(t)\right) \pm O\left(\frac{1}{C^{0.94}}\beta(t)\right) \tag{131}$$

$$= \langle v_i^{(t)}, w_j^{(0)} \rangle \pm \frac{1}{C^{0.9}}\beta(t) \tag{132}$$

Here we use the fact that $\|w_j^{(0)} - w_j^{(t)}\|_2 \leq O\left(\frac{1}{C^{0.95}}\right)$ from the induction hypothesis.

Consider the update of w_i, we have that: at stage 1, we must have $|f(X)|, |f(G(z))| \leq \frac{1}{\text{polylog}(d)}$. Therefore,

$$\nabla_{w_i} L(X, z) = \left(1 \pm \frac{1}{\text{polylog}(d)}\right) a^{(t)}\sigma'(\langle w_i^{(t)}, X\rangle)X - a^{(t)}\sigma'(\langle w_i^{(t)}, G^{(t)}(z)\rangle)G^{(t)}(z) \tag{133}$$

By the induction hypothesis, we have that by $\beta(t) \leq \alpha(t)$, it holds that:

$$\|\sigma'(\langle w_i^{(t)}, G^{(t)}(z)\rangle)G^{(t)}(z)\|_2 \leq O\left(\frac{\alpha(t)}{C} + \frac{1}{C\sqrt{d}}\right)^2 m_G^2 \times m_G\left(\frac{1}{\sqrt{d}} + \frac{\alpha(t)}{C}\right) \leq \epsilon_t$$

On the other hand, we must have that when $X = u_\ell$, we have

$$\left(1 \pm \frac{1}{\text{polylog}(d)}\right)\sigma'(\langle w_{i_\ell^*}^{(t)}, X\rangle) \geq \left(\frac{1}{m_D}\alpha(t) + \frac{1}{\sqrt{d}}\right)^2 \geq \text{polylog}(d)\epsilon_t \tag{134}$$

This completes the proof of the w_i part. For v_i part the proof is the same using the fact that $\|w_j^{(0)} - w_j^{(t)}\|_2 \leq O\left(\frac{1}{C^{0.95}}\right)$ from the induction hypothesis.

□

I.3 STAGE 2 TRAINING

In this stage, we can maintain the following simple update rule: For w_i:

Lemma I.3. *For every $t \in (T_{N,1}, T_1]$, we have that: for every $i \in [m_D]$, for $i = i_\ell^*$:*

$$\mathbb{E}[w_i^{(t+1)}] = w_i^{(t)} + \Theta(\eta_D)u_\ell \pm \eta_D\frac{1}{C^{1.501}} \pm \eta_D\gamma$$

and for $i \neq i_1^*, i_2^*$,

$$\mathbb{E}[w_i^{(t+1)}] = w_i^{(t)} \pm \eta_D\frac{1}{C^{1.501}} \pm \eta_D\gamma$$

For v_i:

$$\mathbb{E}[v_i^{(t+1)}] = v_i^{(t)} + \left(1 \pm \frac{1}{\text{polylog}(d)}\right)\frac{1}{m_G}\eta_G\frac{w_{g_i^*}^{(t)}}{\|w_{g_i^*}^{(t)}\|_2} \pm \eta_G\frac{1}{C^{1.5}} \tag{135}$$

Proof of Lemma I.3. This Lemma can be proved identically to Lemma I.2: By the induction hypothesis, we have

$$|\langle w_i^{(t)}, v_j^{(t)}\rangle| \leq \log^5 \beta(t) \tag{136}$$

Therefore,

$$\|\sigma'(\langle w_i^{(t)}, G^{(t)}(z)\rangle)v_j^{(t)}\|_2 \leq C^{0.01}\beta(t)^3 \leq \frac{1}{C^{1.51}}\alpha(t)^2$$

Which implies that:

$$w_i^{(t+1)} = w_i^{(t)} + \eta_D \frac{\sigma'(\langle w_i^{(t)}, X\rangle)X \pm \frac{1}{C^{1.51}\alpha(t)^2}}{\sqrt{\sum_{j\in[m_D]}(\sigma'(\langle w_j^{(t)}, X\rangle)^2\|X\|_2^2 + \sum_{j\in[m_D]}\left(\sigma'(\langle w_j^{(t)}, X\rangle)^2\right)^3}} \tag{137}$$

Where $\sum_{j\in[m_D]}\left(\sigma'(\langle w_j^{(t)}, X\rangle)^2\right)^3$ comes from the gradient of $a^{(t)}$. By the induction hypothesis we have that $a^{(t)} = \Omega(\alpha(w_{i_\ell^*}, u_\ell, t))$, so we have

$$w_i^{(t+1)} = w_i^{(t)} + \Theta(\eta_D) \frac{\sigma'(\langle w_i^{(t)}, X\rangle)X \pm \frac{1}{C^{1.51}\alpha(t)^2}}{\sqrt{\sum_{j\in[m_D]}(\sigma'(\langle w_j^{(t)}, X\rangle)^2\|X\|_2^2}} \tag{138}$$

On the other hand, by the induction hypothesis, for $\ell \in 2[$: For $i = i_\ell^*$: $\langle w_i^*, u_\ell\rangle \geq \frac{1}{m_D}\alpha(t)$, and for $i \neq i_1^*, i_2^*$: $|\langle w_i^*, X\rangle| \leq O\left(\frac{1}{C}\alpha(t)\right)$.

This implies that: for $i = i_\ell^*$:

$$\mathbb{E}[w_i^{(t+1)}] = w_i^{(t)} + \Theta(\eta_D)u_\ell \pm \eta_D\frac{1}{C^{1.5}} \pm \eta_D\gamma$$

and for $i \neq i_1^*, i_2^*$,

$$\mathbb{E}[w_i^{(t+1)}] = w_i^{(t)} \pm \eta_D\frac{1}{C^{1.5}}\eta_D\gamma$$

Where the additional γ factor comes from the case when $X = u_1 + u_2$ or $X = 0$.

On the other hand, we also know that:

$$\sum_{i\in[m_D]} \sigma'(\langle w_i^{(t)}, v_j^{(t)}\rangle)w_i^{(t)} \tag{139}$$

$$= \sigma'(\langle w_{g_j^*}^{(t)}, v_j^{(t)}\rangle)w_{g_j^*}^{(t)} \pm m_D\left(\frac{1}{C^{0.9}}\right)^2 \langle w_{g_j^*}^{(t)}, v_j^{(t)}\rangle^2 \text{ polyloglogloglog}(d) \tag{140}$$

$$= \sigma'(\langle w_{g_j^*}^{(t)}, v_j^{(t)}\rangle)w_{g_j^*}^{(t)} \pm \sigma'(\langle w_{g_j^*}^{(t)}, v_j^{(t)}\rangle)\frac{1}{C^{1.6}} \tag{141}$$

Notice that $\|w_i^{(t)}\|_2 = \Omega(1)$ so we complete the proof. □

I.4 PROOF OF THE INDUCTION HYPOTHESIS

Now it remains to prove the induction hypothesis:

Stage 1: In this stage, we will use the update Lemma I.2. By the induction hypothesis we know that for $X = u_\ell$,

$$\langle w_j^{(t)}, X\rangle = \alpha(w_j, u_\ell, t) + \alpha(w_j, w_j, 0)\left\langle \frac{w_j^{(0)}}{\|w_j^{(0)}\|_2}, u_\ell\right\rangle \pm O\left(\frac{1}{C^{0.5}\sqrt{d}}\right) \tag{142}$$

This implies that

$$\sum_j \sigma'(\langle w_j^{(t)}, X\rangle)^2\|X\|_2^2 \geq \left(\frac{1}{m_D}\alpha(t) + \frac{1}{\sqrt{d}}\right)^2$$

Now, apply Lemma I.2 we know that:

$$\alpha(t+1) \geq \alpha(t) + \Omega\left(\frac{1}{m_D}\right)\eta_D \tag{143}$$

$$\forall * \neq u_1, u_2 : |\alpha(w_i, *, t+1)| \leq |\alpha(w_i, *, t)| + \eta_D\frac{\epsilon_t}{\left(\frac{1}{m_D}\alpha(t) + \frac{1}{\sqrt{d}}\right)} \leq |\alpha(w_i, *, t)| + \eta_D\frac{1}{C^{1.4}} \tag{144}$$

Compare these two updates we can prove the bounds on w_j for $* \neq u_1, u_2$. For $* = u_1, u_2$, we can see that: By Lemma I.2, there exists $S_{t,\ell} \in (0, \mathrm{poly}(d)]$ such that for $\ell \in [2]$ such that for every $i \in [m_D]$:

$$\sum_{i \in [m_D]} \langle w_i^{(t)}, u_\ell \rangle^4 = \frac{1}{S_{t,\ell}^2} \tag{145}$$

$$\mathbb{E}[\langle w_i^{(t+1)}, u_\ell \rangle] = \langle w_i^{(t)}, u_\ell \rangle + \eta_D \left(1 \pm \frac{1}{\mathrm{polylog}(d)}\right) S_{t,\ell} \langle w_i^{(t)}, u_\ell \rangle^2 \pm \eta_D \frac{1}{\mathrm{polylog}(d)} \tag{146}$$

Apply Lemma E.4 and Lemma I.1 we can complete the proof that

$$|\alpha(w_i, u_1, t)|, |\alpha(w_i, u_2, t)| \leq \alpha(t), \quad \alpha(w_{i_\ell}, u_\ell, t) \geq \frac{\eta_D}{3m_D}$$

and at iteration $t = T_{N,1}$, we have that: for all $i \neq i_1^*, i_2^*$, for all $\ell \in [2]$

$$|\alpha(w_i, u_\ell, t)| \leq \frac{1}{C} \alpha(t)$$

Moreover, when $i = i_{\ell^*}$, $|\alpha(w_i, u_{3-\ell}, t)| \leq \frac{1}{C} \alpha(t)$

The v part can be proved similarly: We have that there exists $S_{t,i} \in (0, \mathrm{poly}(d)]$ where $i \in [m_G]$ such that:

$$\sum_{j \in [m_D]} \langle v_i^{(t)}, w_j^{(0)} \rangle^4 = \frac{1}{S_{t,i}^2} \tag{147}$$

$$\mathbb{E}[\langle v_i^{(t+1)}, w_j^{(0)} \rangle] = \langle v_i^{(t)}, w_j^{(0)} \rangle + \eta_G \frac{1}{m_G} \left(1 \pm \frac{1}{\mathrm{polylog}(d)}\right) S_{t,i} \langle v_i^{(t)}, w_j^{(0)} \rangle^2 \pm \eta_G \frac{\log^5(d)}{C} \tag{148}$$

Apply Lemma E.4 and Lemma I.1, we have that

$$|\alpha(v_i, w_j, t)| \leq \beta(t), \quad \alpha(v_i, w_{g_i^*}, t) \geq \frac{\eta_G}{3m_G^2}$$

. Moreover, at iteration $t = T_{N,1}$, for all $i \in [m_G]$, $j \in [m_D], j \neq g_i^*$:

$$|\langle v_i^{(t)}, w_j^{(0)} \rangle| \leq C^{-0.95} \langle v_i^{(t)}, w_{g_i^*}^{(0)} \rangle$$

Similarly, we can show that for all $* \neq w_j$, $|\alpha(v_i, *, t)| \leq \beta(t)$ and at iteration $t = T_{N,1}$:

$$|\alpha(v_i, *, t)| \leq \frac{1}{C^{0.95}} \beta(t)$$

Using the fact that $\|w_j^{(t)} - w_j^{(0)}\|_2 \leq \frac{1}{C^{0.94}}$

$$\langle v_i^{(t)}, w_j^{(t)} \rangle = \langle v_i^{(t)}, w_j^{(0)} \rangle + \langle v_i^{(t)}, w_j^{(t)} - w_j^{(0)} \rangle = \langle v_i^{(t)}, w_j^{(0)} \rangle \pm \frac{\beta(t)}{C^{0.93}} \tag{149}$$

Notice that $\langle v_i^{(t)}, w_j^{(t)} \rangle \geq \beta(t) C^{-0.01}$ so we show that at iteration $t = T_{N,1}$:

$$|\langle v_i^{(t)}, w_j^{(t)} \rangle| \leq C^{-0.91} \langle v_i^{(t)}, w_{g_i^*}^{(t)} \rangle$$

Similarly, we can show that for every $\ell \in [2]$,

$$|\langle v_i^{(t)}, w_j^{(t)} \rangle| \leq C^{-0.91} \langle v_i^{(t)}, u_\ell \rangle$$

Stage 2: It remains to prove that for all $t \in [T_{N,1}, T_{N,2}]$, we have that

$$|\langle v_i^{(t)}, w_j^{(t)}\rangle| \leq C^{-0.9}\langle v_i^{(t)}, w_{g_i^*}^{(t)}\rangle$$

The rest of the induction hypothesis follows trivially from Lemma I.3. (for the relationship between $a^{(t)}$ and $\alpha(w_{i_\ell^*}, u_\ell, t)$ we can use Lemma E.3).

To prove this, we know that by the update formula:

$$\langle v_i^{(t+1)}, w_j^{(t+1)}\rangle = \langle v_i^{(t+1)}, w_j^{(t)}\rangle + \langle v_i^{(t+1)}, w_j^{(t+1)} - w_j^{(t)}\rangle \tag{150}$$

$$= \langle v_i^{(t)}, w_j^{(t)}\rangle + \langle v_i^{(t+1)} - v_i^{(t)}, w_j^{(t)}\rangle + \langle v_i^{(t+1)}, w_j^{(t+1)} - w_j^{(t)}\rangle \tag{151}$$

Taking expectation, we have that

$$\mathbb{E}[\langle v_i^{(t+1)}, w_j^{(t+1)}\rangle] = \langle v_i^{(t)}, w_j^{(t)}\rangle + \eta_G\left(1 \pm \frac{1}{\text{polylog}(d)}\right)\frac{1}{m_G}\frac{\langle w_{g_i^*}^{(t)}, w_j^{(t)}\rangle}{\|w_{g_i^*}^{(t)}\|_2} \pm \eta_G\frac{1}{C^{1.409}} \tag{152}$$

$$+ \sum_{\ell\in[2]}\frac{\eta_D}{2}\langle v_i^{(t+1)}, u_\ell\rangle 1_{j=i_\ell^*} \pm \eta_D\frac{1}{C^{1.501}} \tag{153}$$

and

$$\mathbb{E}[\langle v_i^{(t+1)}, u_\ell\rangle] = \langle v_i^{(t)}, u_\ell\rangle + \eta_G\left(1 \pm \frac{1}{\text{polylog}(d)}\right)\frac{1}{m_G}\frac{\langle w_{g_i^*}^{(t)}, u_\ell\rangle}{\|w_{g_i^*}^{(t)}\|_2} \pm \eta_G\frac{1}{C^{1.5}} \tag{154}$$

by the induction hypothesis we know that for every $t \leq T_{N,2}$, we have that $\langle w_j^{(t)}, w_{j-1}^{(t)}\rangle \leq \frac{1}{C^{0.95}}$ and $\|w_j^{(t)}\|_2 = [\Omega(1), \text{polyloglogglog}(d)]$, we know that: when $j = g_i^*$

$$\mathbb{E}[\langle v_i^{(t+1)}, w_j^{(t+1)}\rangle] \geq \langle v_i^{(t)}, w_j^{(t)}\rangle + \eta_G\frac{1}{2m_G\,\text{polyloglogglog}(d)} \tag{155}$$

When $j \neq g_i^*$: using the fact that $\eta_G = \eta_D C^{-0.6}$, we have:

$$\mathbb{E}[|\langle v_i^{(t+1)}, w_j^{(t+1)}\rangle|] \leq |\langle v_i^{(t)}, w_j^{(t)}\rangle| + \eta_G\frac{1}{C^{0.9001}} \tag{156}$$

When $i_\ell^* \neq g_i^*$, we have that:

$$\mathbb{E}[|\langle v_i^{(t+1)}, u_\ell\rangle|] \leq |\langle v_i^{(t)}, u_\ell\rangle| + \eta_G\frac{1}{C^{0.95}} \tag{157}$$

Thus we complete the proof.

I.5 Proof of the final theorem

To prove the final theorem, notice that by Lemma I.3, we have that for every $t \in (T_{N,1}, T_1]$, for $i = i_\ell^*$:

$$\mathbb{E}[w_i^{(t+1)}] = w_i^{(t)} + \Theta(\eta_D)u_\ell \pm \eta_D\frac{1}{C^{1.501}} \pm \eta_D\gamma$$

Together with the induction hypothesis, this implies that when $\|w_i^{(t)}\|_2 \geq \log\log\log(d)$, we have that $\langle w_i^{(t)}, u_\ell\rangle \geq (1 - o(1))\|w_i^{(t)}\|_2$. Together with the update formal of $v_j^{(t)}$ we know that when $g_j^* = i_\ell^*$, we have that

$$\mathbb{E}[v_j^{(t+1)}] = v_j^{(t)} + \left(1 \pm \frac{1}{\text{polylog}(d)}\right)\frac{1}{m_G}\eta_G\frac{w_{g_j^*}^{(t)}}{\|w_{g_j^*}^{(t)}\|_2} \pm \eta_G\frac{1}{C^{1.5}} \tag{158}$$

Together with the induction hypothesis, we know that when $\|w_i^{(t)}\|_2 = \text{polyloglog}(d)$, we have that: $\langle v_j^{(t)}, u_\ell\rangle \geq (1 - o(1))\|v_j^{(t)}\|_2$. This proves the theorem.