Journal of Machine Learning Research 26 (2025) 1-61 Submitted 8/23; Revised 12/24; Published 1/25

The Effect of SGD Batch Size on Autoencoder Learning: Sparsity, Sharpness, and Feature Learning

Nikhil Ghosh NIKHIL_GHOSH@BERKELEY.EDU
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720, USA

Spencer Frei SFREI@GOOGLE.COM
Department of Statistics
University of California, Davis
Davis, CA 95616, USA

Wooseok Ha HAYWSE@KAIST.AC.KR
Department of Mathematical Sciences
Korea Advanced Institute of Science & Technology
Daejeon 34141, South Korea

Bin Yu BINYU@STAT.BERKELEY.EDU
Departments of Statistics and EECS
University of California, Berkeley
Berkeley, CA 94720, USA

Editor: Mahdi Soltanolkotabi

Abstract

In this work we investigate the dynamics of stochastic gradient descent (SGD) when training a single-neuron autoencoder with linear or ReLU activation on orthogonal data. We show that for this non-convex problem, randomly initialized SGD with a constant step size successfully finds a global minimum for any batch size choice. However, the particular global minimum found depends upon the batch size. In the full-batch setting, we show that the solution is dense (i.e., not sparse) and is highly aligned with its initialized direction, showing that relatively little feature learning occurs. On the other hand, for any batch size strictly smaller than the number of samples, SGD finds a global minimum which is sparse and nearly orthogonal to its initialization, showing that the randomness of stochastic gradients induces a qualitatively different type of "feature selection" in this setting. Moreover, if we measure the sharpness of the minimum by the trace of the Hessian, the minima found with full batch gradient descent are *flatter* than those found with strictly smaller batch sizes, in contrast to previous works which suggest that large batches lead to sharper minima. To prove convergence of SGD with a constant step size, we introduce a powerful tool from the theory of non-homogeneous random walks which may be of independent interest.

Keywords: Stochastic gradient descent, autoencoder, feature learning, sharpness

1 Introduction

Recent years have witnessed impressive successes of neural networks across a wide variety of domains. However, their ability to generalize to unseen data is still not fully understood

(Zhang et al., 2021; Neyshabur et al., 2017a). One potential explanation is that gradient-based optimization algorithms have an "implicit bias" towards particular solutions which have simple structure, e.g. small norm or rank (Vardi, 2022; Ji and Telgarsky, 2020; Azulay et al., 2021; Gunasekar et al., 2018a,b; Soudry et al., 2018; Neyshabur et al., 2017b; Boursier et al., 2022; Frei et al., 2023b). In some instances, these solutions provably achieve small generalization error (Woodworth et al., 2020; Safran et al., 2022; Frei et al., 2023a). It has been observed that the choice of step size and batch size in these algorithms can make a substantial difference in the generalization performance of trained neural networks, with generally better performance obtained when using larger step sizes and smaller batch sizes (Keskar et al., 2017; Jastrzebski et al., 2017; Wu et al., 2018).

These observations have inspired a surge of research aimed at more deeply understanding the particular effects of small batch size (HaoChen et al., 2021; Damian et al., 2021; Mulayoff and Michaeli, 2020; Pesme et al., 2021; Wu et al., 2022) and step-size (Li et al., 2019; Nacson et al., 2022; Even et al., 2023) on SGD training. However, most prior theoretical works have not directly analyzed the effect of *mini-batch noise*. Instead these works model the gradient noise by considering algorithms which explicitly add noise to the labels or to the gradients, often further assuming a vanishingly small step size (HaoChen et al., 2021; Damian et al., 2021; Xie et al., 2021; Li et al., 2022). An exception is the work by Pesme et al. (2021) who study the effect of mini-batch noise in SGD, but their analysis relies on an infinitesimally small step size. This motivates the main questions we investigate in this work:

Are there settings where ***standard*** *SGD training of a neural network converges to qualitatively different solutions based on the choice of batch size?*

To investigate this question, we consider the setting of a single neuron autoencoder with either a linear or ReLU activation trained on orthonormal data. This neural network model is a simplified version of a prototypical autoencoder where the dynamics of SGD are analytically tractable and reveal a separation between SGD and GD. Orthonormal data can be seen as a special case of a sparse coding model where the dictionary is orthogonal and the latent codes are one-hot vectors (see Section 2 for more details).

To be concrete, we will refer to any training algorithm which updates the parameters by subtracting a multiple of the gradient of a mini-batch as a **mini-batch GD** algorithm. Let m denote the total number of training points. From now on, we will use the following naming conventions

- **Mini-batch SGD** (or just SGD) will refer to the mini-batch GD algorithm where at each iteration a mini-batch of size $b < m$ is drawn uniformly with replacement.
- **Full-batch GD** (or just GD) will refer to the mini-batch GD algorithm where at each iteration the mini-batch is just the full dataset of m points.

Although the model and data we consider are stylized, the setting is rich enough to allow for us to probe the effect of *mini-batch noise* in SGD in a *non-convex* setting, for which there exists a whole *manifold*[1] of global minima. In particular, we show that in this setting there are a number of striking differences in the solutions found by SGD in comparison to

[1] If the data covariance matrix has a unique top eigenvalue then there is a unique global optimum, but if the top eigenvalues repeat, then there is a manifold of global optima (Baldi and Hornik, 1989).

GD. We will say that a solution is "sparse" if it can be expressed as a sparse linear combination of the data, which serves to further highlight the connection with the sparse coding model mentioned earlier. The focus of our results is to highlight an interesting *optimization* phenomenon, rather than to provide insights about *generalization*. Since the data is orthogonal, any test data point orthogonal to the training data will also be orthogonal to the trained neuron and not learned. We now summarize at a high level our main contributions and their implications, which hold for both linear and ReLU activations.

1. In the full-batch setting, randomly initialized GD converges to a *dense* global minimum that is a nearly uniform mixture of many training data points. This minimum is just a rescaling of the initialization projected onto a subset of the span of the data.

2. For any batch size strictly smaller than the size of the dataset, SGD converges almost surely to a single datapoint, which is a 1-*sparse* global minimum that is nearly orthogonal to the random initialization. Notably, the SGD convergence result holds for a constant step size and any batch size strictly smaller than the number of samples.

3. We show that GD exhibits relatively little feature learning since the learned solution is in nearly the same direction as its random initialization, whereas the SGD solution is nearly orthogonal to it. Additionally, the GD solution is invariant to certain orthogonal transformations of the data while the SGD solution is not, further illustrating that SGD learns a more data dependent solution.

4. If we measure the sharpness of the solution found by the trace of the Hessian, we show that SGD converges to *sharper* minima than GD when the activation is ReLU. In contrast, previous works hypothesize that smaller batches result in flatter minima (Keskar et al., 2017), which suggests a potential weakness of this measure of sharpness.

Our results hold by a careful analysis of the trajectory of SGD/GD following a standard random initialization scheme. We show that for orthonormal data, the ReLU autoencoder dynamics reduce to that of a linear autoencoder trained on a subset of the data. Thus it suffices for us to analyse the linear autoencoder dynamics. The loss landscape of linear autoencoders has been studied in the past (Baldi and Hornik, 1989; Plaut, 2018; Kunin et al., 2019), as well as their gradient dynamics (Gidel et al., 2019; Bao et al., 2020) which are closely related to Oja's rule from neuroscience (Oja, 1982; Yan et al., 1994) and the streaming PCA problem (Shamir, 2016; Allen-Zhu and Li, 2017). However, no prior work has studied the convergence of gradient methods in parameter space when the first principal component is not unique, as is the case in our setting; nor has prior work highlighted the role of batch size. In particular, the SGD case requires significant technical innovation as we are considering the dynamics under a constant (fixed) step size and we cannot couple the SGD trajectory with that of GD since they converge to qualitatively different minima.

Indeed, most classical analyses of SGD (e.g., Robbins and Monro (1951)) assume the step size decays to zero as this is what is generally required to ensure that the iterates can converge to single point rather than to a measure with full support. However, it is not always required to decay the step size for training to converge. Notably, Nacson et al. (2019) proved that for linearly separable data, a linear model trained with SGD on the logistic loss with a constant step size converges in direction to the ℓ_2-max-margin predictor.

This however is identical to the convergence behavior of full-batch GD (Soudry et al., 2018), hence it is impossible to isolate the effect of stochastic gradients on the types of solutions found by SGD/GD in this setting. Pesme et al. (2021) investigates the different generalization behaviors of SGD and GD, but for an infinitesimally small step size. In practice, constant step size SGD often suffices to fully optimize the training loss and achieve competitive generalization capabilities (Soudry et al., 2018), making it a practical baseline as it requires less tuning than more complicated schedules. In particular, step size decay is usually employed not to better optimize the training loss, but to improve the generalization performance.

To analyze constant step size SGD in our setting we introduce a powerful tool from the theory of non-homogeneous random walks to develop convergence guarantees. Such tools first appeared in the probability literature (Menshikov and Wade, 2010), but to our knowledge have not been employed in a machine learning setting prior to this work. At a high level, our proof works by showing that at each iteration SGD amplifies the correlation of the weights with a particular data direction relative to all other data directions. By viewing this relative correlation as a stochastic process induced by SGD, we can invoke our tool to show that the stochastic process is transient and then show that this implies convergence to a global minimum. In contrast, full-batch GD exhibits a symmetry which ensures that this process remains fixed at initialization. This symmetry is broken by the random subsampling of SGD which leads to the “phase transition” in the asymptotic convergence behavior when the batch size changes. Note that although the limiting convergence behavior of SGD is equivalent for all batch sizes, different sizes can yield different “rates of escape” of the associated stochastic process, which leads to different asymptotic convergence rates. We believe the techniques we develop for the convergence of constant step size SGD in this setting may hold wider applicability for the analysis of other machine learning algorithms.

The rest of the paper is organized as follows. In Section 2, we provide a review of related work, and Section 3 formally presents our problem setting. We then present our theoretical results for the linear autoencoder in Section 4, where we provide proof sketches for the convergence of GD and SGD in Section 4.1 and Section 4.5, respectively. In Section 5, we present the corresponding convergence results for the ReLU autoencoder, and in Section 5.1 we compare the local loss landscapes of the ReLU autoencoder at different algorithmic solutions. Finally, we conclude in Section 6 where we also provide potential avenues for further work.

2 Related Work

Linear Autoencoders and Streaming PCA. It has long been known that there is a strong connection between PCA and linear neural networks. Baldi and Hornik (1989) showed that a linear autoencoder with squared loss has a minimum which is the projection of the data onto the subspace spanned by the first principal components of the training data. A variety of works actually analyze algorithms for recovering the PCA subspace. Oja (1982) proposed a biologically plausible update rule known as Oja’s rule for training a single neuron in an online setting to recover the first PCA direction. The computer science community has considered other algorithms including variations of Oja’s rule for solving PCA in the streaming setting in a space efficient manner (Shamir, 2016; Allen-Zhu and

Li, 2017; Mitliagkas et al., 2013; Jain et al., 2016). There have also been works which specifically analyze PCA recovery via training linear autoencoders using gradient descent or gradient flow (Gidel et al., 2019; Min et al., 2021; Saxe et al., 2013). However, all of these works only consider convergence of the loss to its minimum and not the convergence of the weights which we show can be heavily dependent on the choice of batch size. It is important to study the learned weights as we do in this work since in practice these correspond to learned features which can have a significant impact when used for downstream tasks.

Neural networks and sparsity. The sparse coding data model (i.e., data of the form $x = Az + \varepsilon$, where the "latent code" z is sparse, the "dictionary" A is unknown, and ε is a noise variable) is a widely-used generative model for natural data (Olshausen and Field, 1997; Vinje and Gallant, 2000; Olshausen and Field, 2004). A number of previous works have studied different "dictionary learning" algorithms designed to recover the hidden dictionary A given data generated from the sparse coding model e.g., Arora et al. (2015); Agarwal et al. (2016). More closely related to our setting is Nguyen et al. (2019) which shows that two-layer autoencoders trained by variants of full-batch gradient descent on sparse coding data can *locally* converge to the ground truth dictionary. In our work, we also analyze gradient descent trained autoencoders trained on a (simplified instance of) sparse coding data. However, we provide *global* convergence guarantees across a range of batch sizes.

Batch size, sharpness, and generalization. It has been empirically observed that in practice large batch size tends to degrade the generalization performance of SGD (Keskar et al., 2017) on supervised learning tasks. One contending hypothesis for this behavior is that larger batch sizes result in SGD finding "sharper" minima which generalize poorly. Recently, there have been several theoretical works which try to make this intuition rigorous by studying SGD with explicitly added label noise (Blanc et al., 2020; Damian et al., 2021; HaoChen et al., 2021; Li et al., 2022). At a high level these works show that the added label noise has the following implicit regularization effect: once the iterates reach a global minimum of the training loss the iterates approximately remain on the manifold of global minima, but now move to decrease a regularization term. This regularization term can be viewed as the sharpness of the loss and is approximately equal to the trace of the Hessian for small step sizes. One can show that for certain problems such as sparse overparametrized linear regression (Woodworth et al., 2020) that decreasing this notion of sharpness leads to better generalization. We show that this notion of sharpness may not be universally appropriate, since in our setting smaller batch size leads to a *sharper* solution. Moreover, in comparison to prior works (Blanc et al., 2020; Damian et al., 2021; HaoChen et al., 2021; Li et al., 2022), we do not require independent noise to be added to the gradient updates to model SGD noise but rather we explicitly characterize the effect of the randomness that comes from using stochastic gradients in SGD. Pesme et al. (2021) also consider the randomness from mini batches in SGD, but their analysis is limited to the continuous version of SGD.

Feature learning in neural networks. Neural networks trained by gradient descent have shown a remarkable ability to learn data-dependent features which enable generalization to a variety of domains. A number of recent works have explored how different aspects of the training procedure, including network architecture and optimization hyperparameters, affect this 'feature learning' ability. Jacot et al. (2018); Chizat et al. (2019) showed that when the width of neural networks grows to infinity, the learning rate is small and the random initialization has a large variance, the training dynamics of the network can be approximated by the behavior of a data-independent kernel defined by an infinite-width limit of the network at its random initialization. In this 'kernel regime' setting, the network behaves similarly to a *random* feature model and no data-dependent feature learning occurs. By contrast, when the learning rate is large and the scale of initialization is small, neural networks are indeed capable of learning data-dependent features, as has been shown in a number of recent works (Wei et al., 2019; Allen-Zhu and Li, 2022; Yang and Hu, 2021; Frei et al., 2022; Zou et al., 2023). However, none of these works examined how the batch size of SGD could significantly affect the types of features found by SGD, as we do in this work. In our work, we see that despite the fact that the weights move non-trivially for both GD and SGD, it turns out that SGD displays more data-dependent feature learning behavior.

Dynamics of GD for single-neurons. We note that previous works have also considered the dynamics of gradient descent for learning single-neuron architectures, e.g., Yehudai and Shamir (2020); Frei et al. (2020); Vardi et al. (2021); Mei et al. (2018). However, none of these previous works considered unsupervised learning with autoencoders or establish a separation between the minima learned using gradient descent with different batch sizes.

Convergence of SGD with a fixed step-size. Notably, we guarantee that the final iterate of constant step-size SGD converges almost surely to a single point. Typical convergence guarantees for SGD require either a decaying step-size, iterate averaging, or only hold in expectation or with high probability (Jain et al., 2019; Sebbouh et al., 2021; Pesme et al., 2021; Liu and Yuan, 2022; Zou et al., 2021). Prior works have shown that for constant step size SGD the (averaged) iterates almost surely converge to an invariant distribution (Merad and Gaïffas, 2023; Yu et al., 2020). In general the limiting invariant distribution will have non-zero variance, however our problem has the interesting feature that the last iterate converges to a single point mass with zero variance. To the best of our knowledge, the only other such example is provided in Nacson et al. (2019).

3 Setting

We consider a single neuron weight-tied auto-encoder $f : \mathbb{R}^n \to \mathbb{R}^n$ defined as

$$f(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{w}\phi(\langle \boldsymbol{w}, \boldsymbol{x} \rangle). \tag{1}$$

The network takes as input $\boldsymbol{x} \in \mathbb{R}^n$ and is parameterized by a single neuron $\boldsymbol{w} \in \mathbb{R}^n$ with activation ϕ and no bias. We will take the activation ϕ to be either the identity $\phi(z) = z$ or ReLU $\phi(z) = \max(0, z)$. Furthermore, we will assume that we are given a training dataset $\mathcal{D} = \{\boldsymbol{a}_1, \dots, \boldsymbol{a}_m\}$ where the $\boldsymbol{a}_i \in \mathbb{R}^n$ are orthonormal and necessarily $m \leq n$. Let $(\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_n)$ be the completion of $\mathcal{D}$ to an orthonormal basis of $\mathbb{R}^n$. We will be interested in characterizing the dynamics of (stochastic) gradient descent on the standard reconstruction objective

$$\mathcal{L}(\boldsymbol{w}; \mathcal{D}) = \frac{1}{m} \sum_{i=1}^{m} \ell(\boldsymbol{w}; \boldsymbol{a}_i), \tag{2}$$

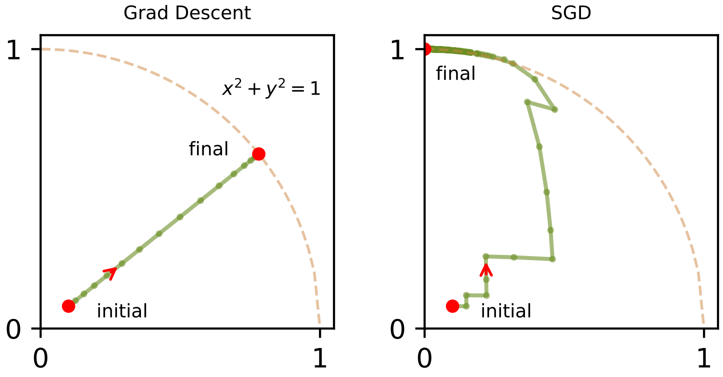

Figure 1: Visualizations of the trajectories of $\boldsymbol{w}_t \in \mathbb{R}^2$ for GD and SGD (with $b = 1$) when $m = n = 2$. Both methods are initialized at $\boldsymbol{w}_0 = (0.1, 0.08)^\top$ and run with step size $\alpha = 1/4$ on the dataset $\mathcal{D} = \{\boldsymbol{a}_1, \boldsymbol{a}_2\}$ consisting of the standard basis vectors.

where the pointwise loss ℓ is the squared-loss

$$\ell(\boldsymbol{w}; \boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x} - f(\boldsymbol{x}; \boldsymbol{w})\|^2.$$

We will consider mini-batch GD training with non-zero initialization and constant step-size $\eta > 0$, namely for $t = 0, 1, \ldots$

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \sum_{i \in \mathcal{B}_t} \boldsymbol{\nabla}_{\boldsymbol{w}}\ \ell(\boldsymbol{w}; \boldsymbol{a}_i), \quad \mathcal{B}_t \subseteq [m], \tag{3}$$

where the gradient of the pointwise loss $\boldsymbol{\nabla}_{\boldsymbol{w}}\ \ell(\boldsymbol{w}; \boldsymbol{x})$ is

$$\phi'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle) \cdot [\boldsymbol{x}\boldsymbol{w}^\top + \langle \boldsymbol{w}, \boldsymbol{x}\rangle \cdot \mathbf{I}_n] \cdot (f(\boldsymbol{x}; \boldsymbol{w}) - \boldsymbol{x}), \tag{4}$$

and we take $\phi'(t) := \mathbb{1}(t > 0)$ when ϕ is ReLU.

There are many possible instantiations of mini-batch GD training based on the mini-batch selection in Eq. (3). We will consider algorithms with fixed batch size $b := |\mathcal{B}_t|$ where $b \in [m]$. In particular, we consider the following algorithms:

- Full-batch GD where $\mathcal{B}_t = [m]$ for all t. Note that in this case $b = m$.
- Mini-batch SGD where each $\mathcal{B}_t$ is chosen uniformly at random from the set of subsets of $[m]$ of size b and $b < m$.
- Cyclic SGD[2] where $\mathcal{B}_t = \{t \bmod m\}$. Note that in this case $b = 1$.

We will often just refer to mini-batch SGD as SGD and full-batch GD as GD for short.

3.1 Visualizations of Convergence Behavior

To demonstrate how the batch size influences the solutions found by gradient descent, we train a linear autoencoder using full-batch GD and stochastic GD on an example where

[2] Note that the mini-batch order is actually deterministic.

$\mathcal{D} = \{\boldsymbol{a}_1, \boldsymbol{a}_2\}$ with $\boldsymbol{a}_1 = (1,0)^\top$ and $\boldsymbol{a}_2 = (0,1)^\top$. Both methods are initialized at the same point $\boldsymbol{w}_0 = (0.1, 0.08)^\top$. Since the iterates $\boldsymbol{w}_t$ lie in $\mathbb{R}^2$, we can visualize the optimization trajectories for each method in Figure 1. In the figure, we also draw the upper quadrant of the unit circle as a dashed curve. Later in Section 5.1 we will show that all points on the quarter circle are global minima. We see that full-batch GD converges to a point in the interior of the quarter circle, whereas SGD converges to a boundary point. As we will see in the next section, we can theoretically understand the behavior of these simulations.

4 Theoretical Results for Linear Activation

In this section we will consider the convergence behavior of the single neuron linear autoencoder define in Eq. (1) when trained with full-batch GD, mini-batch SGD, and cyclic SGD. We will start by stating the results and then in later sections provide proof sketches of GD and SGD convergence, as well as additional relevant background. More specifically, in Section 4.1 we give a proof sketch of GD convergence. In the remaining sections we work towards sketching the proof of SGD convergence. We start with giving some useful results about the iterates of mini-batch (S)GD in Section 4.2 that are mostly algebraic. In Section 4.3 we identify a stochastic process arising from SGD which is key to understanding its convergence behavior. To analyze this process, in Section 4.4 we introduce a useful probabilistic result from the theory of non-homogeneous random walks. Finally in Section 4.5 we combine all the previous results to give a proof sketch of mini-batch SGD convergence.

Let us start by introducing some notation. Given a set of indices $\mathcal{S} \subseteq [n]$, we write the orthogonal projection onto $\operatorname{span}(\boldsymbol{a}_i : i \in \mathcal{S})$ as $\Pi_{\mathcal{S}}$ where

$$\Pi_{\mathcal{S}}(\boldsymbol{x}) := \sum_{i \in \mathcal{S}} \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle \boldsymbol{a}_i. \tag{5}$$

For convenience, define $\Pi_m := \Pi_{[m]}$. We now present our convergence result for full-batch GD. We give a sketch of the proof in Section 4.1 and present the full proof in Appendix D.

Theorem 1 (GD) *Assume that $\|\boldsymbol{w}_0\| < 1$ and $0 < \eta \leq 1/5$. Then the full-batch GD iterates $\boldsymbol{w}_t$ converge to the point $\boldsymbol{w}_{\mathrm{GD}}$ as $t \to \infty$ where*

$$\boldsymbol{w}_{\mathrm{GD}} := \frac{\Pi_m(\boldsymbol{w}_0)}{\|\Pi_m(\boldsymbol{w}_0)\|},$$

and the projection Π_m is defined in Eq. (5).

Our result states that full-batch GD converges to the point obtained by taking the initialization $\boldsymbol{w}_0$, projecting it onto the span of the data, and then rescaling it to have norm one. Note that this implies that GD is invariant to orthogonal transformations of the data which preserve the span of the dataset. In Section 5.1, we will show that $\boldsymbol{w}_{\mathrm{GD}}$ is a global minimum of the objective function.

We will now give the convergence behavior of mini-batch SGD. First define the set of initially positive datapoint directions

$$\mathcal{S}^+ := \{i \in [m] : \langle \boldsymbol{w}_0, \boldsymbol{a}_i \rangle > 0\}. \tag{6}$$

The following theorem provides a limiting characterization for mini-batch SGD. A sketch of the proof is given in Section 4.5 and the full proof is given in Appendix C.

Theorem 2 (SGD) *Assume $\|\boldsymbol{w}_0\| < 1$ and $0 < \eta \leq 1/5$. Then the mini-batch SGD iterates $\boldsymbol{w}_t$ converge to an element of the set $\{\boldsymbol{a}_i : i \in \mathcal{S}^+\} \cup \{-\boldsymbol{a}_j : j \in [m] \setminus \mathcal{S}^+\}$ a.s. over the randomness of SGD minibatch sampling.*

Note that our theorem applies for a deterministic choice of initialization $\boldsymbol{w}_0$ and dataset $\mathcal{D}$ and holds almost surely over the sampling of the mini-batches (recall that at each step of SGD, a batch is sampled uniformly at random from the distinct batches of size $b < m$). Crucially, the convergence behavior of SGD does not depend on the step size, as long as the batch size is strictly smaller than the full batch. Our result does not give the precise probability distribution over the m possible limit points, however we are still able to infer that almost surely $\boldsymbol{w}_t$ eventually converges to a single point and perfectly aligns itself with some element of the dataset. Convergence to a point with a constant step-size is possible due to the fact that all of the pointwise gradients vanish at the SGD limit points. Like $\boldsymbol{w}_{\text{GD}}$, the solutions found by SGD are also global minima of the objective function; we shall show this in Section 5.1.

For random initialization $\boldsymbol{w}_0 \sim \mathcal{N}(0, (\sigma_{\text{init}}^2/n) \cdot \mathbf{I}_n)$, Theorem 1 implies that the GD solution $\boldsymbol{w}_{\text{GD}}$ is a random unit vector in $\text{span}(\boldsymbol{a}_i : i \in [m])$. If m is large then the iterates will hardly correlate with any particular data direction, but will by highly correlated with the initialization $\boldsymbol{w}_0$. This is in contrast to the SGD solution $\boldsymbol{w}_{\text{SGD}}$ which by Theorem 2 is perfectly correlated with some datapoint and nearly orthogonal to the intialization. To make this quantitative, let us define for $\boldsymbol{w} \in \mathbb{R}^n$ its cosine similarity with $\boldsymbol{x} \in \mathbb{R}^n$ which we denote as $\text{cossim}(\boldsymbol{w}, \boldsymbol{x})$ and its maximum cosine similarity with the dataset $\mathcal{D} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m\}$ which we denote as $\text{cossim}(\boldsymbol{w}, \mathcal{D})$, as follows

$$\text{cossim}(\boldsymbol{w}, \boldsymbol{x}) := \left| \left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \right\rangle \right|, \quad \text{cossim}(\boldsymbol{w}, \mathcal{D}) := \max_{\boldsymbol{x} \in \mathcal{D}} \text{cossim}(\boldsymbol{w}, \boldsymbol{x}). \tag{7}$$

We have the following corollary which gives the limiting cosine similarities with the dataset and the initialization for GD and SGD with random initialization.

Corollary 3 *Assume that $\boldsymbol{w}_0 \sim \mathcal{N}(0, (\sigma_{\text{init}}^2/n) \cdot \mathbf{I}_n)$ where $\sigma_{\text{init}} < 1$. If $0 < \eta \leq 1/5$, then with probability at least $1 - O(m^{-1})$, the GD iterates $\boldsymbol{w}_t^{\text{GD}}$ satisfy*

$$\lim_{t \to \infty} \text{cossim}(\boldsymbol{w}_t^{\text{GD}}, \mathcal{D}) = O\left(\sqrt{\frac{\log m}{m}}\right), \quad \lim_{t \to \infty} \text{cossim}(\boldsymbol{w}_t^{\text{GD}}, \boldsymbol{w}_0) = \Theta\left(\frac{m}{n}\right),$$

and the iterates of minibatch SGD satisfy

$$\lim_{t \to \infty} \text{cossim}(\boldsymbol{w}_t^{\text{SGD}}, \mathcal{D}) = 1, \quad \lim_{t \to \infty} \text{cossim}(\boldsymbol{w}_t^{\text{SGD}}, \boldsymbol{w}_0) = O\left(\sqrt{\frac{\log m}{n}}\right),$$

where $\text{cossim}(\boldsymbol{w}_t, \mathcal{D})$ is defined in Eq. (7).

Corollary 3 illustrates that when n is large and m is on the same order as n, with high probability SGD finds solutions which are significantly different to its random initialization, while the GD solution is quite close to its random initialization in cosine similarity. Furthermore, the GD solution is invariant to orthogonal transformations which preserve the span of the data, which roughly means that the solution only depends on the data through the linear subspace spanned by the data. All together, these observations support the view that SGD exhibits a stronger form of data-dependent feature learning than GD in this setting. [3]

One may wonder if the convergence behavior of minibatch SGD in our setting is primarily driven by the stochasticity of the mini-batch selection. In the next theorem, we suggest this may not be the case and that the more important property is the batch-size. Namely, we show that a totally deterministic cyclic selection of mini-batch indices still leads to convergence to a single datapoint when the iterates are initialized from a certain non-trivial region with positive Lebesgue measure.

Theorem 4 (CSGD) *Let $m = n = 2$ and $\mathcal{D} = \{\boldsymbol{a}_0, \boldsymbol{a}_1\}$. Assume that $\langle \boldsymbol{w}_0, \boldsymbol{a}_0 \rangle \geq \langle \boldsymbol{w}_0, \boldsymbol{a}_1 \rangle$ and $\langle \boldsymbol{w}_0, \boldsymbol{a}_1 \rangle > 0$. Furthermore, assume that $\|\boldsymbol{w}_0\| < 1$, and $0 < \eta \leq 1/4$. Then the CSGD iterates $\boldsymbol{w}_t$ converge to $\boldsymbol{a}_0$ as $t \to \infty$.*

We provide the proof of this result in Appendix E. Unlike our mini-batch SGD result, here we are able to explicitly determine the convergence point and give support to the intuition that the updates are biased towards converging to the $\boldsymbol{a}_i$ the iterate is currently maximally correlated with. Compared to the mini-batch SGD result, however, this result is limited by the fact that we restrict to the two-dimensional setting $m = n = 2$ and impose the initialization condition $\langle \boldsymbol{w}_0, \boldsymbol{a}_0 \rangle \geq \langle \boldsymbol{w}_0, \boldsymbol{a}_1 \rangle$. We believe the result should hold more broadly for arbitrary m, n and $\boldsymbol{w}_0$ such that $\|\boldsymbol{w}_0\| < 1$ and we have empirically verified this behavior in simulations. However, the proof for CSGD is quite complicated even in the $m = n = 2$ case, and we focus instead on SGD as our proof for this setting captures arbitrary m and n and we believe holds promise for generalizing to other problem settings.

4.1 Full-batch GD Proof Sketch

In this section we give a proof sketch for Theorem 1. Since the $\boldsymbol{a}_i$ are orthonormal we can analyse the evolution of $\boldsymbol{w}_t$ in terms of the coordinates $c_t(i) = \langle \boldsymbol{w}_t, \boldsymbol{a}_i \rangle$. The detailed proof is given in Appendix D. One can write the full-batch gradient descent updates of each coordinate $c_t(i)$ as follows,

$$\begin{aligned} c_{t+1}(i) &= c_t(i)(1 + \eta(2 - 2\Phi_t - \Psi_t)), && i \in [m] \\ c_{t+1}(j) &= c_t(j)(1 - \eta\Phi_t), && j \in [n] \setminus [m] \end{aligned}$$

where we define the quantities

$$\Phi_t := \sum_{i \in [m]} c_t(i)^2, \quad \Psi_t := \sum_{j \in [n] \setminus [m]} c_t(j)^2. \tag{8}$$

It is easy to see that for full-batch GD, the ratio of the coordinate updates $c_{t+1}(i)/c_t(i)$ is the same for each coordinate i. Using this, we can derive the following key invariant,

$$c_t(i) = c_0(i)\sqrt{\frac{\Phi_t}{\Phi_0}}, \quad \text{for all } i \in [m] \text{ and all } t \in \{0, 1, \ldots\}. \tag{9}$$

[3] We are not aware of an agreed-upon definition of feature learning but a common view in the deep learning literature is that more feature learning occurs if the solution found is far from its initialization and incorporates more data-dependent information (Chizat et al., 2019; Woodworth et al., 2020).

Thus, in order to understand the dynamics of $c_t(i)$ for $i \in [m]$, it suffices to understand the dynamics of Φ_t. Similarly, by understanding the dynamics of Ψ_t we can characterize the dynamics of $c_t(j)$ for $j \in [n] \setminus [m]$. It turns out that $(\Phi_{t+1}, \Psi_{t+1}) \in \mathbb{R}^2$ can be written solely in terms of $(\Phi_t, \Psi_t) \in \mathbb{R}^2$, so we instead directly analyze the evolution of this two-dimensional system. Under the conditions $\|\boldsymbol{w}_0\| < 1$ and step-size $\eta \leq 1/5$, we can establish the following boundedness property for all t,

$$\Phi_t + (5/8)\Psi_t < 1, \text{ for all } t \in \{0, 1, \ldots\}. \tag{10}$$

Using Eq. (10), it is not hard to show $\Phi_t \to 1$ and $\Psi_t \to 0$ as $t \to \infty$. Now the result follows since $c_t(i) \to c_0/\sqrt{\Phi_0}$ for $i \in [m]$ by Eq. (9) and $c_t(j) \to 0$ for $i \notin [m]$.

4.2 Properties of Mini-batch (S)GD

We now move on to sketching the proof of mini-batch SGD convergence. We start with some general properties of mini-batch (S)GD which hold for any mini-batch sequence with batch size $b < m$. The proofs of the results in this section can be found in Appendices A and B. In particular, the proofs of Proposition 5 and Corollary 6 can be found in Appendix A and the proof of Proposition 7 can be found in Appendix B. As before we will analyze the evolution of the coordinates $c_t(i)$. From Eq. (3) and Eq. (4) we have,

$$\begin{aligned} c_{t+1}(i) &= c_t(i)\Big(1 + \eta\Big(2 - \|\Pi_{\mathcal{B}_t}(\boldsymbol{w}_t)\|^2 - \|\boldsymbol{w}_t\|^2\Big)\Big) && i \in \mathcal{B}_t, \\ c_{t+1}(j) &= c_t(j)\Big(1 - \eta\|\Pi_{\mathcal{B}_t}(\boldsymbol{w}_t)\|^2\Big) && j \notin \mathcal{B}_t. \end{aligned}$$

The next result states that for any mini-batch sequence the iterates are bounded in ℓ_2-norm.

Proposition 5 (Bounded Iterates) *Assume that $\|\boldsymbol{w}_0\| < 1$ and $0 < \eta \leq 1/5$. Then for all $t \geq 0$ and any batch size $b < m$, the iterates of mini-batch GD for any mini-batch sequence $(\mathcal{B}_t)_{t\geq 0}$ satisfy*

$$\|\boldsymbol{w}_t\|^2 \leq 1 + \eta/4.$$

Note that although this bound is weaker than the corresponding one in Eq. (10) for full-batch GD, it still provides several useful consequences. One can show that $c_{t+1}(\ell)/c_t(\ell) > 0$ for all $\ell \in [n]$. Thus the coordinates never change sign. Moreover for $j \notin \mathcal{B}_t$, $c_{t+1}(j)/c_t(j) < 1$, hence the magnitude of coordinates not present in the batch decreases each iteration, that is $|c_{t+1}(j)| < |c_t(j)|$. In particular, Ψ_t as defined in Eq. (8) is decreasing with t. Lastly, one can show that the coordinate magnitudes $|c_t(\ell)| < 1$ which is sharper than the ℓ_2-norm bound from Proposition 5. We summarize these conclusions in the following corollary.

Corollary 6 *Under the conditions of Proposition 5, for all times t we have $|c_t(i)| < 1$ and $\operatorname{sign}(c_t(i)) = \operatorname{sign}(c_0(i))$ for all $i \in [n]$. Furthermore, Ψ_t is monotonically decreasing.*

So far from Corollary 6 we know that for any mini-batch sequence that Ψ_t is decreasing and from 5 that $\|\boldsymbol{w}_t\|$ is bounded. The next result shows that under the additional assumption that the mini-batch sequence is chosen at random, we can say more.

Proposition 7 *For mini-batch SGD with any batch size $b < m$ the following hold,*

1. *As $t \to \infty$, almost surely $\Psi_t \to 0$,*
2. *Almost surely* $\liminf_{t\to\infty} \|\boldsymbol{w}_t\| \geq 1$.

Since Ψ_t as defined in Eq. (8) is orthogonal projection onto the complement of the data subspace, the above essentially states that eventually the iterates lie in the subspace spanned by the data and have norm at least one. We will now start to more specifically describe our proof strategy for SGD convergence. The proof relies on connecting the convergence of SGD with a particular stochastic process which we describe in the next section.

4.3 SGD and Random Walks

The connection between SGD and random walks[4] arises since, roughly speaking, at each step the neuron $\boldsymbol{w}_t$ "rotates" in the direction of $\boldsymbol{a}_i$ in the mini-batch with which it has the highest correlation. We wish to show that asymptotically $\boldsymbol{w}_t$ becomes completely aligned with one point $\boldsymbol{a}_i$ and completely unaligned with all other datapoints. We can track this relative alignment by analysing a certain one-dimensional random walk which is a function of the iterates. Let $i_t^\star = \arg\max_{i\in[m]} |c_t(i)|$ be the direction with highest alignment and $\mathcal{J}_t = [m] \setminus \{i_t^\star\}$ be the set of remaining directions. The random walk we analyse is the log-ratio quantity $\{R_t\}_{t=0}^{\infty}$ where

$$R_t := \log\left(\frac{|c_t(i_t^\star)|}{\sum\limits_{\ell\in\mathcal{J}_t} |c_t(\ell)|}\right). \tag{11}$$

Our goal is to show that $R_t \to \infty$ almost surely, as this will imply that the neuron is completely aligned with some direction in the limit as $t \to \infty$. Conveniently, we only need to consider a single ratio involving the most aligned direction $i_t^\star$ since we are not concerned with which particular $\boldsymbol{a}_i$ the iterates converge to. In the next section, we describe tools for analysing stochastic processes of this type.

4.4 Non-homogeneous Random Walks

We now present a result from the theory of non-homogeneous random walks that is used in our proof of Theorem 2. We remark that this theory is much broader in scope than what we present and has been used before to analyse other stochastic systems such as urn processes, birth-and-death chains, etc. However, to the best of our knowledge we present the first application of this theory to the analysis of SGD. We believe that such techniques should be broadly useful in analysing the behavior of SGD for other problems.

We now introduce the notation and assumptions. Define $\mathbb{Z}^+ := \{0, 1, \ldots\}$ and let $X = (X_t)_{t\in\mathbb{Z}^+}$ be a discrete time stochastic process adapted to a filtration $(\mathcal{F}_t)_{t\in\mathbb{Z}^+}$ and taking values in the half-line $[b_0, \infty)$ for some $b_0 \in \mathbb{R}$. Let us state the basic assumptions.

(A1) For any $y \in (b_0, \infty)$ there exists a function $v : \mathbb{Z}^+ \to \mathbb{Z}^+$ and $\varepsilon > 0$ such that

$$\inf_{t\in\mathbb{Z}^+} \Pr\big[X_{t+v(t)} > y \mid \mathcal{F}_t\big] > \varepsilon, \ \textit{a.s.}$$

[4] We use the term random walk more generally than to just mean a stochastic process arising from a sequence of partial sums of i.i.d random variables.

(A2) For some $K < \infty$

$$\sup_{t \in \mathbb{Z}^+} |X_{t+1} - X_t| \leq K, \text{ a.s.}$$

The first condition is a type of fairly weak irreducibility condition which states that at any time there is at least some positive probability to exceed a value y after some number of steps and implies in particular that $\limsup_{t \to \infty} X_t = \infty$ a.s. The second condition states the process has bounded increments. Under these conditions we have the following result which is a special case of the more general Theorem 2.2 from Menshikov and Wade (2010) and can be used to show the process X is transient.

Proposition 8 (Menshikov and Wade (2010)) *Let $X = (X_t)_{t \in \mathbb{Z}^+}$ be a stochastic process adapted to the filtration $(\mathcal{F}_t)_{t \in \mathbb{Z}^+}$ on the half-line $[b_0, \infty)$ for some $b_0 \in \mathbb{R}$. Assume that Assumptions (A1) and (A2) hold for X. If there exists a function $\underline{\mu}_1 : [b_0, \infty) \to \mathbb{R}$ such that for all $t \in \mathbb{Z}^+$*

$$\underline{\mu}_1(X_t) \leq \mathbb{E}[X_{t+1} - X_t \mid \mathcal{F}_t], \text{ a.s.},$$

and $\liminf_{x \to \infty} \underline{\mu}_1(x) > 0$, then X is transient, that is $X_t \to \infty$ a.s. as $t \to \infty$.

4.5 SGD Proof Sketch

Using the previous results we can now prove Theorem 2. Complete proofs for this section can be found in Appendix C. In the following proposition we establish the transience of the process $R = (R_t)_{t \in \mathbb{Z}^+}$ by verifying that it obeys the conditions of Proposition 8. Note that it makes sense to apply this result since

$$R_t = \log \left(\frac{|c_t(i_t^\star)|}{\sum\limits_{\ell \in \mathcal{J}_t} |c_t(\ell)|} \right) \geq \log \left(\frac{|c_t(i_t^\star)|}{|\mathcal{J}_t||c_t(i_t^\star)|} \right) = -\log(|\mathcal{J}_t|) = -\log(m-1),$$

hence R is a stochastic process on $[b_0, \infty)$ where $b_0 = -\log(m-1)$.

Proposition 9 (Transience of R) *The process $(R_t)_{t \in \mathbb{Z}^+}$ defined in Eq. (11) arising from SGD is transient, i.e., $R_t \to \infty$ a.s. as $t \to \infty$.*

Let us first try to gain some intuition about this result. Note that for full-batch GD we actually have that $R_t = R_0$ for all t, hence this result is not true for full-batch GD. This is because for full-batch GD $c_{t+1}(i)/c_t(i) = 1 + \eta(2 - 2\Phi_t - \Psi_t)$ for all $i \in [m]$, hence $R_{t+1} = R_t$. Now let us try to see why the process is transient for mini-batch SGD.

Sketch of Proposition 9 To invoke Proposition 8 we will need to analyze the increment of the process $R_{t+1} - R_t$. More specifically, we will need to show that the conditional expected increment is lower bounded by a positive constant, in addition to verifying Assumptions (A1) and (A2). Let us first observe that the increment $R_{t+1} - R_t \geq \Delta_t$ where

$$\Delta_t := \log \left(\frac{|c_{t+1}(i_t^\star)|}{\sum_{\ell \in \mathcal{J}_t} |c_{t+1}(\ell)|} \right) - \log \left(\frac{|c_t(i_t^\star)|}{\sum_{\ell \in \mathcal{J}_t} |c_t(\ell)|} \right). \tag{12}$$

This follows since $|c_{t+1}(i^\star_{t+1})| \geq |c_{t+1}(i^\star_t)|$ by definition and

$$\sum_{\ell \in \mathcal{J}_{t+1}} |c_{t+1}(\ell)| = \sum_{\ell \in [m]} |c_{t+1}(\ell)| - |c_{t+1}(i^\star_{t+1})| \leq \sum_{\ell \in [m]} |c_{t+1}(\ell)| - |c_{t+1}(i^\star_t)| = \sum_{\ell \in \mathcal{J}_t} |c_{t+1}(\ell)|,$$

hence the first term on the right of Eq. (12) is less than R_{t+1} and the second term is just R_t. Thus to lower bound the increment, it will suffice to lower bound the more tractable quantity Δ_t. Intuitively, we should expect that Δ_t will be positive when $i^\star_t \in \mathcal{B}_t$ and negative otherwise. By considering these two cases we will show that in expectation Δ_t is positive. Importantly however, note that from the statement of Proposition 8 that we only need to establish this lower bound *asymptotically* for large R_t and large times t. For the purposes of the proof sketch we will informally use the notation $A \gtrapprox B$ to denote that the inequality is true up to an error which vanishes as $R_t \to \infty$ and $t \to \infty$, with $A \approx B$ taken to mean $A \gtrapprox B$ and $A \lessapprox B$. In particular, we have the following asymptotic statement

$$\|\boldsymbol{w}_t\|^2 \approx c_t(i^\star_t)^2 \approx 1. \tag{13}$$

To see why this holds note that since $|c_t(\ell)| < 1$ by Corollary 6,

$$\|\boldsymbol{w}_t\|^2 - c_t(i^\star_t)^2 - \Psi_t = \sum_{\ell \in \mathcal{J}_t} c_t(\ell)^2 \leq \sum_{\ell \in \mathcal{J}_t} |c_t(\ell)| = |c_t(i^\star_t)| \exp(-R_t) \leq \exp(-R_t).$$

By Proposition 7, $\Psi_t \to 0$, hence from the above $\|\boldsymbol{w}_t\|^2 \approx c_t(i^\star_t)^2$. Also, $\liminf \|\boldsymbol{w}_t\|^2 \geq 1$ hence $\|\boldsymbol{w}_t\|^2 \gtrapprox 1$. Since $c_t(i^\star_t)^2 \leq 1$ this yields Eq. (13).

Now we move on to lower bounding $\mathbb{E}(\Delta_t \mid \mathcal{F}_t)$ by a quantity which becomes a time-independent positive constant as $R_t \to \infty$ and $t \to \infty$. For simplicity, in this sketch let us consider the setting where $b = 1$ and $c_0(\ell) > 0$ for all $\ell \in [n]$. By Corollary 6, $c_t(\ell) \in (0, 1)$ for all t. Let $i_t \in [m]$ be the selected mini-batch index. With probability $1/m$, we have $i_t = i^\star_t$. In this case one can calculate that

$$\Delta_t = \log\left(\frac{1 + \eta(2 - c_t(i^\star_t)^2 - \|\boldsymbol{w}_t\|^2)}{1 - \eta c_t(i^\star_t)^2}\right) \approx \log\left(\frac{1}{1-\eta}\right) \tag{14}$$

where we used that $\|\boldsymbol{w}_t\|^2 \approx c_t(i^\star_t)^2 \approx 1$. On the other hand, if $i_t = i$ for some $i \neq i^\star_t$ then

$$-\Delta_t = \log\left(\frac{1}{1 - \eta c_t(i_t)^2} \frac{\sum_{\ell \in \mathcal{J}_t} c_{t+1}(\ell)}{\sum_{\ell \in \mathcal{J}_t} c_t(\ell)}\right)$$

where we can expand the term

$$\begin{aligned}\frac{\sum_{\ell \in \mathcal{J}_t} c_{t+1}(\ell)}{\sum_{\ell \in \mathcal{J}_t} c_t(\ell)} &= \frac{\eta c_t(i_t)(1 + \eta(2 - c_t(i_t)^2 - \|\boldsymbol{w}_t\|^2)) + (1 - \eta c_t(i_t)^2) \sum_{\ell \in \mathcal{J}_t \setminus \{i_t\}} c_t(\ell)}{\sum_{\ell \in \mathcal{J}_t} c_t(\ell)} \\ &= \frac{c_t(i_t)(2 - \|\boldsymbol{w}_t\|^2) + (1 - \eta c_t(i_t)^2) \sum_{\ell \in \mathcal{J}_t} c_t(\ell)}{\sum_{\ell \in \mathcal{J}_t} c_t(\ell)} \\ &\approx \frac{\eta c_t(i_t)}{\sum_{\ell \in \mathcal{J}_t} c_t(\ell)} + (1 - \eta c_t(i_t)^2).\end{aligned}$$

Combining with the earlier expression and using that $c_t(i_t)^2 \approx 0$ by Eq. (13) gives

$$\begin{aligned}
-\Delta_t &\approx \log\left(1 + \frac{1}{1-\eta c_t(i_t)^2}\frac{\eta c_t(i_t)}{\sum_{\ell\in\mathcal{J}_t} c_t(\ell)}\right) \\
&\approx \log\left(1 + \frac{\eta c_t(i_t)}{\sum_{\ell\in\mathcal{J}_t} c_t(\ell)}\right) \\
&\le \eta \cdot \frac{c_t(i_t)}{\sum_{\ell\in\mathcal{J}_t} c_t(\ell)}.
\end{aligned}$$

Therefore combining with Eq. (14) we have that conditioned on $\mathcal{F}_t$

$$\begin{aligned}
\mathbb{E}\Delta_t &\gtrapprox \Pr(i_t = i_t^\star)\log\left(\frac{1}{1-\eta}\right) - \sum_{i\in\mathcal{J}_t}\Pr(i_t = i)\cdot\eta\cdot\frac{c_t(i)}{\sum_{\ell\in\mathcal{J}_t} c_t(\ell)} \\
&= \frac{1}{m}\left[\log\left(\frac{1}{1-\eta}\right) - \eta\right] \ge \frac{\eta^2}{2m},
\end{aligned}$$

where the last inequality uses $\log(1/(1-x)) \ge x(1+x/2)$ for $x \in (0,1)$. This accomplishes our goal of asymptotically bounded the expected conditional increment. It is not much harder to verify Assumptions (A1) and (A2). To see that (A1) holds, note that if $i_t = i_t^\star$, then from Eq. (14), using that $\|\boldsymbol{w}_t\|^2 \le 1 + \eta/4$ from Proposition 5 yields

$$\begin{aligned}
\Delta_t &= \log\left(\frac{1 + \eta(2 - c_t(i_t^\star)^2 - \|\boldsymbol{w}_t\|^2)}{1 - \eta c_t(i_t^\star)^2}\right) \\
&= \log\left(1 + \eta\frac{2 - \|\boldsymbol{w}_t\|^2}{1 - \eta c_t(i_t^\star)^2}\right) \\
&\ge \log(1 + \eta(1 - \eta/4)) > 0,
\end{aligned}$$

where the last inequality holds since $\eta(1-\eta/4) > 0$. Thus on this event we have lower bounded the increment by a time-independent constant. Furthermore the probability of this event is $1/m$ which is also time-independent. Therefore we can see that Assumption (A1) will be satisfied by considering the event that $i_t = i_t^\star$ a sufficiently large (but time-independent) number of times in a row. Verifying Assumption (A2) is not very difficult and should be plausible given that $|c_t(\ell)| < 1$ and $\|\boldsymbol{w}_t\|^2 \le 1 + \eta/4$.

Sketch of Theorem 2 Now taking Proposition 9 to be true, the rest of the proof follows quite easily. By Corollary 6,

$$\begin{aligned}
1 \ge c_t(i_t^\star)^2 &= \|\boldsymbol{w}_t\|^2 - \sum_{\ell\in\mathcal{J}_t} c_t(\ell)^2 - \Psi_t \\
&\ge \|\boldsymbol{w}_t\|^2 - \exp(-R_t) - \Psi_t.
\end{aligned}$$

Now by Propositions 7 and 9, we have $\liminf \|\boldsymbol{w}_t\| \ge 1$, $\Psi_t \to 0$, and $R_t \to \infty$ as $t \to \infty$, so we see that $|c_t(i_t^\star)| \to 1$. One can then show that $i_t^\star$ must eventually become constant since the gradient norm goes to zero when $\boldsymbol{w}_t$ approaches any $\boldsymbol{a}_i$. Therefore there exists some $i^\star \in [m]$ such that $i_t^\star = i^\star$ eventually. By Corollary 6 we have that $\operatorname{sign}(c_t(i^\star)) = \operatorname{sign}(c_0(i^\star))$, hence $\boldsymbol{w}_t \to \operatorname{sign}(c_0(i^\star)) \cdot \boldsymbol{a}_i$ as we wished to show.

5 Theoretical Results for ReLU Activation

In this section, we will consider the case when the activation function of the autoencoder in Eq. (1) is the ReLU $\phi(t) = \max(t, 0)$.[5] Our results will rely upon an equivalence between the dynamics of SGD/GD for ReLU autoencoders with the dynamics of SGD/GD for linear autoencoders. To this end, let us introduce some preliminary notation. Let us denote the output and losses for the ReLU autoencoder and linear autoencoder as,

$$\begin{aligned} &f(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{w} \max(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, 0) \qquad && \ell(\boldsymbol{w}; \boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x} - f(\boldsymbol{x}; \boldsymbol{w})\|^2 \qquad && \text{(ReLU autoencoder)},\\ &\widetilde{f}(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{w}\langle \boldsymbol{w}, \boldsymbol{x} \rangle && \widetilde{\ell}(\boldsymbol{w}; \boldsymbol{x}) = \frac{1}{2}\left\|\boldsymbol{x} - \widetilde{f}(\boldsymbol{x}; \boldsymbol{w})\right\|^2 && \text{(Linear autoencoder)}. \end{aligned}$$

Let

$$\mathcal{S}_t^+ = \{i \in [m] : \langle \boldsymbol{a}_i, \boldsymbol{w}_t \rangle > 0\}, \qquad \widetilde{\mathcal{B}}_t := \mathcal{B}_t \cap \mathcal{S}_t^+.$$

The set $\mathcal{S}_t^+$ consists of datapoints which the neuron is positively correlated with at time t, and $\widetilde{\mathcal{B}}_t$ is the subset of samples in the batch selected at time t which are also in $\mathcal{S}_t^+$. Our key observation is that the minibatch GD update for the ReLU autoencoder with minibatch $\mathcal{B}_t$ is equivalent to the minibatch GD update if the activation was linear and the minibatch was instead the set $\widetilde{\mathcal{B}}_t = \mathcal{B}_t \cap \mathcal{S}_t^+$. That is, the update in Eq. (3) satisfies

$$\boldsymbol{w}_t - \eta \sum_{i \in \mathcal{B}_t} \boldsymbol{\nabla}_{\boldsymbol{w}}\ \ell(\boldsymbol{w}_t; \boldsymbol{a}_i) = \boldsymbol{w}_t - \eta \sum_{i \in \widetilde{\mathcal{B}}_t} \boldsymbol{\nabla}_{\boldsymbol{w}}\ \widetilde{\ell}(\boldsymbol{w}_t; \boldsymbol{a}_i), \tag{15}$$

simply since for $i \in [m]$,

$$\boldsymbol{\nabla}_{\boldsymbol{w}}\ \ell(\boldsymbol{w}; \boldsymbol{a}_i) = \begin{cases} \boldsymbol{\nabla}_{\boldsymbol{w}}\ \widetilde{\ell}(\boldsymbol{w}; \boldsymbol{a}_i) & \text{if } i \in \mathcal{S}_t^+, \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

By Corollary 6 in Section 4.2, if we define $\mathcal{S}^+ := \mathcal{S}_0^+$ then we know that if $\|\boldsymbol{w}_0\| < 1$ and $\eta \leq 1/5$, then for any minibatch GD algorithm $\mathcal{S}_t^+ = \mathcal{S}^+$ for all t and so $\widetilde{\mathcal{B}}_t = \mathcal{B}_t \cap \mathcal{S}^+$.

For full-batch GD, $\widetilde{\mathcal{B}}_t = [m] \cap \mathcal{S}^+ = \mathcal{S}^+$, hence full-batch GD on a ReLU autoencoder with initialization $\boldsymbol{w}_0$ is equivalent to running full-batch GD on a linear autoencoder with initialization $\boldsymbol{w}_0$ but with the dataset $\mathcal{S}^+$ instead of $[m]$. Thus from Theorem 1 it is easy to see that we have the following theorem for GD convergence for a ReLU autoencoder.

Theorem 10 (GD-ReLU) *Assume that $\|\boldsymbol{w}_0\| < 1$ and $\eta \leq 1/5$. Define the set $\mathcal{S}^+$ as in Eq. (6). Then the full-batch GD iterates $\boldsymbol{w}_t$ converge to the point $\boldsymbol{w}_{\mathrm{GD}}$ as $t \to \infty$ where*

$$\boldsymbol{w}_{\mathrm{GD}} := \frac{\Pi_{\mathcal{S}^+}(\boldsymbol{w}_0)}{\|\Pi_{\mathcal{S}^+}(\boldsymbol{w}_0)\|}.$$

For the case of mini-batch SGD, note that $|\widetilde{\mathcal{B}}_t| \leq |\mathcal{B}_t| < m$. If $|\widetilde{\mathcal{B}}_t| = 0$, then nothing happens that iteration, and we can just focus on the subsequence where $|\widetilde{\mathcal{B}}_t| > 0$. By Eq.

[5] Note that the ReLU is non-differentiable at the origin, although a sub-gradient exists for $\phi(t)$ at every $t \in \mathbb{R}$. The only issue that could arise is if the weights are exactly orthogonal to one of the $\boldsymbol{a}_i$, but with a Gaussian random initialization this does not occur almost surely.

(15) we see that minibatch SGD on the ReLU autoencoder is equivalent to a minibatch GD algorithm on a linear autoencoder with dataset $\mathcal{S}^+$ where the minibatch $\widetilde{\mathcal{B}}_t \subseteq \mathcal{S}^+$ has a random batch-size that can vary with time. More specifically, we can view the process of selecting $\widetilde{\mathcal{B}}_t$ as first randomly choosing the effective batch size $\tilde{b}_t := |\widetilde{\mathcal{B}}_t| \in \{1, \ldots, b\}$, and then conditioned on this choice $\widetilde{\mathcal{B}}_t$ is chosen uniformly from the set of subsets of $\mathcal{S}^+$ of size $\tilde{b}_t$, that is, $\widetilde{\mathcal{B}}_t$ is a batch size $\tilde{b}_t$ minibatch SGD selection from $\mathcal{S}^+$.

With the above in mind, we can essentially transfer the minibatch SGD proof for the linear case to the ReLU setting by viewing $\mathcal{S}^+$ as the effective dataset. Accordingly, one can show that the stochastic process

$$\widetilde{R}_t := \log\left(\frac{c_t(i_t^\star)}{\sum\limits_{j \in \mathcal{S}^+ \setminus \{i_t^\star\}} c_t(j)}\right), \quad i_t^\star := \operatorname*{arg\,max}_{i \in \mathcal{S}^+} c_t(i),$$

which arises from replacing $[m]$ with $\mathcal{S}^+$ in the definition of R_t in Eq. (22) is transient. Note that the proof of the transience of the stochastic process R in Proposition 9 relied only on showing that the properties of the increment $R_{t+1} - R_t$ required by Proposition 8 hold, which was done for any batch size $b < m$ (see Appendix C.1). Thus these properties hold conditionally on $\tilde{b}_t$, from which it is not hard to see that they extend to hold unconditionally as well, hence the process $\widetilde{R}_t$ is indeed transient. From there, following the same logic as in the rest of the proof of Theorem 2 (see Appendix C.2), it is easy to see that the following holds for ReLU autoencoder.

Theorem 11 (SGD-ReLU) *Assume that $\|\boldsymbol{w}_0\| < 1$ and $0 < \eta \leq 1/5$. Then the minibatch SGD iterates $\boldsymbol{w}_t$ converge to some element of the set $\{\boldsymbol{a}_i : i \in \mathcal{S}^+\}$ almost surely.*

5.1 Loss Landscape

In this section we will study properties of the loss landscape of the ReLU autoencoder at the points which GD and SGD converge to. Our first result characterizes the set of global minima of the loss objective.

Theorem 12 (Global Minima) *The minimum value of the loss objective $\mathcal{L}(\boldsymbol{w})$ from Eq. (2) is attained on*

$$\mathcal{M} = \left\{ \sum_{i=1}^m c_i \boldsymbol{a}_i : c_1, \ldots, c_m \geq 0 \text{ and } \sum_{i=1}^m c_i^2 = 1 \right\},$$

where it achieves the value $(m-1)/(2m)$.

A visualization of the loss landscape is given in Figure 2 when $m = n = 2$. We note that a similar argument used to prove Theorem 12 shows that for the case of the linear autoencoder, the global minima are attained on the set $\widetilde{\mathcal{M}} = \{\sum_{i=1}^m c_i \boldsymbol{a}_i : \sum_{i=1}^m c_i^2 = 1\}$, which is defined just like $\mathcal{M}$ except the coefficients c_i are not required to be non-negative.

By the result above and our convergence theorems we can see that both GD and SGD converge to global minima. Indeed, Theorem 10 shows that full batch gradient descent converges to the following solution

$$\boldsymbol{w}_{\mathrm{GD}} = \sum_{i \in S} \frac{\langle \boldsymbol{w}_0, \boldsymbol{a}_i \rangle}{\sqrt{\Phi}} \boldsymbol{a}_i, \qquad \Phi = \sum_{i \in \mathcal{S}^+} \langle \boldsymbol{w}_0, \boldsymbol{a}_i \rangle^2 \tag{16}$$

where $\mathcal{S}^+$ is defined in Eq. (6). By the above theorem $\boldsymbol{w}_{\mathrm{GD}}$ is a global minimum. From Theorem 11 we see that SGD converges to

$$\boldsymbol{w}_{\mathrm{SGD}} = \boldsymbol{a}_i, \quad \text{for some } i \in \mathcal{S}^+. \tag{17}$$

Again by Theorem 12 this point is also a global minimum. Thus, these algorithms optimally minimize the loss objective, but achieve qualitatively different solutions. For random intializations, SGD learns a “pure” datapoint whereas GD learns a “mixture”.

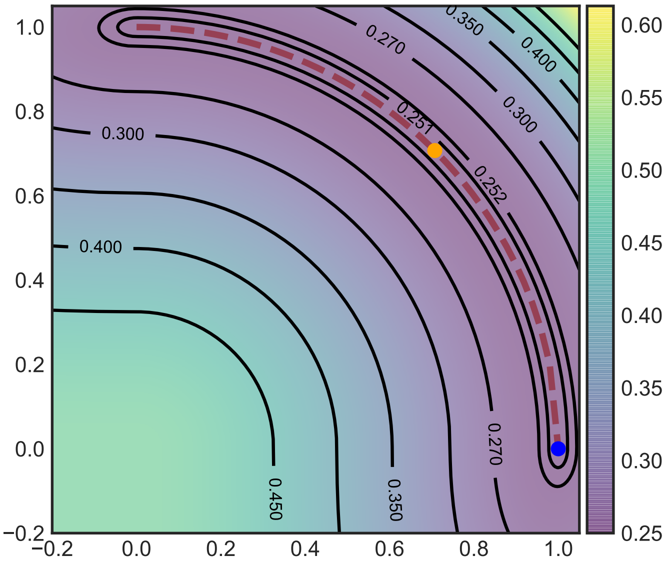

Figure 2: Contour plot of the loss objective $\mathcal{L}(\boldsymbol{w})$ where $m = n = 2$ and $\boldsymbol{a}_1 = (1, 0)^\top$ and $\boldsymbol{a}_2 = (0, 1)^\top$. The red dotted line shows the set of global minima $\mathcal{M}$ given in Theorem 12. The blue circle shows one possible solution that can be obtained with SGD ($\boldsymbol{w}_{\mathrm{SGD}} = \boldsymbol{a}_1$), whereas the orange circle shows a solution that can be obtained with GD but not with SGD ($\boldsymbol{w}_{\mathrm{GD}} = 1/\sqrt{2} \cdot \boldsymbol{a}_1 + 1/\sqrt{2} \cdot \boldsymbol{a}_2$).

In recent years, much of the literature on deep learning has sought to distinguish the learned solutions of large and small batch GD by the “sharpness” of the obtained solution. The prevailing intuition is that small batch sizes lead to flatter minima (which is correlated with better generalization). However, we show that for our problem, this is not true for common measures of sharpness which either fail to distinguish the two types of solutions or lead to the conclusion that smaller batches lead to *sharper* minima. Hence, the claim that smaller batch sizes leads to sharper minima is not true without adding further assumptions or adjusting the definition of sharpness.

The measures of sharpness we consider come from the eigenspectrum of the Hessian. Namely, if $\boldsymbol{H}$ is the Hessian of the loss at a given point, then we consider either the maximal eigenvalue $\|\boldsymbol{H}\|_2$ or the sum of the eigenvalues $\mathrm{Tr}(\boldsymbol{H})$, where larger values indicate sharper points. The following result gives the sharpness at the points $\boldsymbol{w}_{\mathrm{GD}}$ and $\boldsymbol{w}_{\mathrm{SGD}}$.

Theorem 13 *Denote the Hessians*[6] *of the loss $\mathcal{L}(\boldsymbol{w})$ at the points $\boldsymbol{w}_{\mathrm{GD}}$ and $\boldsymbol{w}_{\mathrm{SGD}}$ defined in Eq. (16), (17) as $\boldsymbol{H}_{\mathrm{GD}}$ and $\boldsymbol{H}_{\mathrm{SGD}}$ respectively. Then,*

$$\|\boldsymbol{H}_{\mathrm{GD}}\|_2 = \frac{4}{m}, \quad \mathrm{Tr}(\boldsymbol{H}_{\mathrm{GD}}) = \frac{2n+8-m-|\mathcal{S}^+|}{2m},$$
$$\|\boldsymbol{H}_{\mathrm{SGD}}\|_2 = \frac{4}{m}, \quad \mathrm{Tr}(\boldsymbol{H}_{\mathrm{SGD}}) = \frac{2n+7-m}{2m}.$$

From the above we have the following observations. Both $\boldsymbol{H}_{\mathrm{GD}}$ and $\boldsymbol{H}_{\mathrm{SGD}}$ have the same maximum eigenvalue, hence this measure fails to distinguish $\boldsymbol{w}_{\mathrm{GD}}$ and $\boldsymbol{w}_{\mathrm{SGD}}$. For large m, if $\boldsymbol{w}_0$ is initialized by a random Gaussian then with high probability $|\mathcal{S}^+| \approx m/2$, hence

$$\mathrm{Tr}(\boldsymbol{H}_{\mathrm{GD}}) = \frac{2n+8-m-|\mathcal{S}^+|}{2m} \approx \frac{2n+8-m-m/2}{2m}$$

from which it follows

$$\mathrm{Tr}(\boldsymbol{H}_{\mathrm{GD}}) \approx \mathrm{Tr}(\boldsymbol{H}_{\mathrm{SGD}}) - \frac{1}{2} < \mathrm{Tr}(\boldsymbol{H}_{\mathrm{SGD}}),$$

hence $\boldsymbol{w}_{\mathrm{SGD}}$ is *sharper* than $\boldsymbol{w}_{\mathrm{GD}}$ as claimed.

6 Conclusion

In this work, we investigated the dynamics of SGD and GD for training single neuron autoencoders with ReLU activation following random initialization on orthogonal data. We showed that for any choice of batch size, SGD/GD converge to global minima despite nonconvexity. However, the particular minimum found depends strongly upon the batch size. In the full-batch deterministic setting, GD converges to a minimum which is dense (i.e., not sparse) with respect to the training data and is highly aligned with its initial direction. As such, relatively little feature-learning occurs in the full-batch setting. For any batch size strictly smaller than the number of samples, SGD converges to a sparse global minimum which is almost orthogonal to its initialization, hence SGD exhibits stronger feature learning compared to GD. Moreover, SGD finds solutions which are sharper than those found by full-batch GD if we measure sharpness by the trace of the Hessian. We are able to prove that the SGD iterates converge almot surely to a degenerate point mass distribution even with a constant step size by introducing and using tools from the literature on non-homogeneous random walks. We believe that these tools may be more broadly applicable for the analysis of other machine learning algorithms.

It is worth emphasizing that although we find that the minima found by SGD can be distinguished from those found by GD by sharpness, the relationship we find (SGD

[6] As explained in Appendix F, due to the non-differentiability of the ReLU function, we consider more general measures based on one-sided derivatives which extend the Hessian based definitions.

producing sharper minima) is the opposite of that found by prior work (Keskar et al., 2017). This suggests that sharpness may not be the most useful way to characterize the effect of batch size in the dynamics of SGD. On the other hand, we also show that at least for single neuron autoencoders, the effect of batch size can be distinguished through the lens of sparsity and through the lens of feature learning: we find that SGD prefers sparse minima and that stochastic gradients lead to significantly different features than those found at random initialization.

There are a number of natural directions for future work. One interesting direction is to understand whether the effects of batch size on the dynamics of SGD/GD we find in this work extend to more complex models and distributions, especially for classification problems. Moreover, because we work with orthogonal data, there is not a well-defined notion of "generalization". We thus were unable to investigate the effect of batch size on generalization, which is an important question.

Another interesting direction is to see if the tools we developed for the almost-sure convergence of SGD with a finite step size can be used in other settings. Our technique (see Proposition 8) relied upon the work by Menshikov and Wade (2010) and required constructing a particular stochastic process where the transience of this process implies convergence of SGD, reducing the convergence of SGD to the transience of a stochastic process. The choice of this stochastic process leveraged insights into the structure of the problem, including the interplay between data, loss function, and SGD dynamics. We believe a similar approach can work in other settings. Moreover, although we did not pursue this line of investigation due to space limitations, it should be possible to develop rates of the convergence of this stochastic process by using other results by Menshikov and Wade (2010).

Acknowledgments and Disclosure of Funding

This research is kindly supported in part by the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and 814639. NG would like to acknowledge support from the NSF RTG Grant #1745640, and BY would like to acknowledge partial support from the NSF Grant DMS-2015341. WH was supported from NSF TRIPODS Grant 1740855, DMS-1613002, and 2015341.

References

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, and Praneeth Netrapalli. Learning sparsely used overcomplete dictionaries via alternating minimization. *SIAM Journal on Optimization*, 26(4):2775–2799, 2016.

Zeyuan Allen-Zhu and Yuanzhi Li. First efficient convergence for streaming k-pca: a global, gap-free, and near-optimal rate. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 487–492. IEEE, 2017.

Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, 2022.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. In *Conference on learning theory*, pages 113–149. PMLR, 2015.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *International Conference on Machine Learning*, pages 468–477. PMLR, 2021.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.

Xuchan Bao, James Lucas, Sushant Sachdeva, and Roger B Grosse. Regularized linear autoencoders recover the principal components, eventually. *Advances in Neural Information Processing Systems*, 33:6971–6981, 2020.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR, 2020.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. *arXiv preprint arXiv:2206.00939*, 2022.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (s) gd over diagonal linear networks: Implicit regularisation, large stepsizes and edge of stability. *arXiv preprint arXiv:2302.08982*, 2023.

Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of a single neuron with gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Spencer Frei, Niladri S. Chatterji, and Peter L. Bartlett. Random feature amplification: Feature learning and generalization in neural networks. *Preprint, arXiv:2202.2202.07626*, 2022.

Spencer Frei, Gal Vardi, Peter L. Bartlett, and Nathan Srebro. The double-edged sword of implicit bias: Generalization vs. robustness in relu networks. *Preprint, arXiv:2303.01456*, 2023a.

Spencer Frei, Gal Vardi, Peter L. Bartlett, Nathan Srebro, and Wei Hu. Implicit bias in leaky ReLU networks trained on high-dimensional data. In *International Conference on Learning Representations*, 2023b.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in Neural Information Processing Systems*, 31, 2018b.

Jeff Z HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Conference on Learning Theory*, pages 2315–2357. PMLR, 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Prateek Jain, Chi Jin, Sham M Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming pca: Matching matrix bernstein and near-optimal finite sample guarantees for oja's algorithm. In *Conference on learning theory*, pages 1147–1164. PMLR, 2016.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of sgd information theoretically optimal. In *Conference on Learning Theory (COLT)*, 2019.

Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

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

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations (ICLR)*, 2017.

Daniel Kunin, Jonathan Bloom, Aleksandrina Goeva, and Cotton Seed. Loss landscapes of regularized linear autoencoders. In *International conference on machine learning*, pages 3560–3569. PMLR, 2019.

Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Neural Information Processing Systems*, 2019.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? –a mathematical framework. In *International Conference on Learning Representations*, 2022.

Jun Liu and Ye Yuan. On almost sure convergence rates of stochastic gradient methods. In *Conference on Learning Theory (COLT)*, 2022.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

Mikhail V Menshikov and Andrew R Wade. Rate of escape and central limit theorem for the supercritical lamperti problem. *Stochastic processes and their applications*, 120(10): 2078–2099, 2010.

Ibrahim Merad and Stéphane Gaïffas. Convergence and concentration properties of constant step-size sgd through markov chains. *arXiv preprint arXiv:2306.11497*, 2023.

Hancheng Min, Salma Tarmoun, René Vidal, and Enrique Mallada. On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In *International Conference on Machine Learning*, pages 7760–7768. PMLR, 2021.

Ioannis Mitliagkas, Constantine Caramanis, and Prateek Jain. Memory limited, streaming pca. *Advances in neural information processing systems*, 26, 2013.

Rotem Mulayoff and Tomer Michaeli. Unique properties of flat minima in deep networks. In *International Conference on Machine Learning*, pages 7108–7118. PMLR, 2020.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit bias of the step size in linear diagonal neural networks. In *International Conference on Machine Learning*, pages 16270–16295. PMLR, 2022.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017a.

Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017b.

Thanh V Nguyen, Raymond KW Wong, and Chinmay Hegde. On the dynamics of gradient descent for autoencoders. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2858–2867. PMLR, 2019.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15:267–273, 1982.

Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325, 1997.

Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487, 2004. ISSN 0959-4388.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

Elad Plaut. From principal subspaces to principal components with linear autoencoders. *arXiv preprint arXiv:1804.10253*, 2018.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Itay Safran, Gal Vardi, and Jason D. Lee. On the effective number of linear regions in shallow univariate reLU networks: Convergence guarantees and implicit bias. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Othmane Sebbouh, Robert M. Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory (COLT)*, 2021.

Ohad Shamir. Convergence of stochastic gradient descent for pca. In *International Conference on Machine Learning*, pages 257–265. PMLR, 2016.

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

Gal Vardi. On the implicit bias in deep-learning algorithms. *arXiv preprint arXiv:2208.12591*, 2022.

Gal Vardi, Gilad Yehudai, and Ohad Shamir. Learning a single neuron with bias using gradient descent. *Advances in Neural Information Processing Systems*, 34:28690–28700, 2021.

William E. Vinje and Jack L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287 5456:1273–6, 2000.

Colin Wei, Jason D. Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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

Lei Wu, Chao Ma, et al. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

Lei Wu, Mingze Wang, and Weijie Su. When does sgd favor flat minima? a quantitative characterization via linear stability. *arXiv preprint arXiv:2207.02628*, 2022.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2021.

Wei-Yong Yan, Uwe Helmke, and John B Moore. Global analysis of oja's flow for neural networks. *IEEE Transactions on Neural Networks*, 5(5):674–683, 1994.

Greg Yang and Edward J. Hu. Feature learning in infinite-width neural networks. In *International Conference on Machine Learning (ICML)*, 2021.

Gilad Yehudai and Ohad Shamir. Learning a single neuron with gradient methods. In *Conference on Learning Theory*, pages 3756–3786. PMLR, 2020.

Lu Yu, Krishnakumar Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. An analysis of constant step size sgd in the non-convex regime: Asymptotic normality and bias. *arXiv preprint arXiv:2006.07904*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Benign overfitting of constant-stepsize sgd for linear regression. In *Conference on Learning Theory*, pages 4633–4635. PMLR, 2021.

Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. Understanding the generalization of adam in learning neural networks with proper regularization. In *International Conference on Learning Representations (ICLR)*, 2023.

In the appendix we provide missing proofs from the main paper and some additional background on non-homogeneous random walk theory. In appendix sections A, B, C, D, and E we provide convergence results for the linear autoencoder. In Appendix A we give general results concerning the iterates of minibatch GD for any minibatch sequence and in Appendix B we give similar results under the additional assumption that the minibatch indices are chosen randomly. Then in Appendix B we prove the SGD convergence result Theorem 2, in Appendix D we prove the GD convergence result Theorem 2, and in Appendix E we prove the CSGD convergence result Theorem 4. Finally in Appendix F we provide proofs for results about the loss landscape of the ReLU autoencoder.

Appendix A. General Properties of Minibatch Gradient Descent

In this section we will present some general properties of the iterates of minibatch gradient descent on a linear autoencoder (i.e., $\phi(t) = t$ in Eq. (1)) for an arbitrary sequence of minibatch indices, which will be useful for later sections. First let us establish some notation. For a set $\mathcal{S} \subseteq [n]$ define the orthogonal projection onto the directions in $\mathcal{S}$ as

$$\Pi_{\mathcal{S}}(\boldsymbol{x}) := \sum_{i \in \mathcal{S}} \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle \boldsymbol{x},$$

with $\Pi_m := \Pi_{[m]}$. Define the coordinate $c_i(t) := \langle \boldsymbol{w}_t, \boldsymbol{a}_i \rangle$ for $i \in [n]$. We will define

$$\Phi_t := \|\Pi_m(\boldsymbol{x})\|^2 = \sum_{i \in [m]} c_t(i)^2, \tag{18}$$

$$\Psi_t := \|\boldsymbol{w}_t\|^2 - \|\Pi_m(\boldsymbol{x})\|^2 = \sum_{j \in [n] \setminus [m]} c_t(j)^2. \tag{19}$$

Let us write the minibatch GD update in these coordinates. We shall repeatedly use this formulation in the remaining proofs. The updates are given in the following lemma.

Lemma 14 (Coordinate Updates) *Consider the minibatch GD updates given in Eq. (3). If we define $c_t(\ell) = \langle \boldsymbol{w}_t, \boldsymbol{a}_\ell \rangle$ for $\ell \in [n]$, then the updates are equivalently given by*

$$\begin{aligned} c_{t+1}(i) &= c_t(i)(1 + \eta(2 - u_t - \|\boldsymbol{w}_t\|^2)) && i \in \mathcal{B}_t, \\ c_{t+1}(j) &= c_t(j)(1 - \eta u_t) && j \in [n] \setminus \mathcal{B}_t, \end{aligned}$$

where we define $u_t := \|\Pi_{\mathcal{B}_t}(\boldsymbol{w}_t)\|^2$.

Proof We would like to derive equations defining $c_{t+1}(k) = \langle \boldsymbol{w}_{t+1}, \boldsymbol{a}_k \rangle$ for each k. This will depend upon whether or not $k \in \mathcal{B}_t$. Using the definition of the minibatch GD updates in Eq. (3), for any i, $k \in [m]$ we have,

$$\begin{aligned} \langle \nabla \ell(\boldsymbol{w}; \boldsymbol{a}_i), \boldsymbol{a}_k \rangle &= \langle \boldsymbol{w} \boldsymbol{a}_i \boldsymbol{w}^\top + \langle \boldsymbol{w}, \boldsymbol{a}_i \rangle \mathbf{I}_n (\boldsymbol{w} \langle \boldsymbol{w}, \boldsymbol{a}_i \rangle - \boldsymbol{a}_i), \boldsymbol{a}_k \rangle \\ &= \langle \boldsymbol{a}_i, \boldsymbol{a}_k \rangle \|\boldsymbol{w}\|^2 \langle \boldsymbol{w}, \boldsymbol{a}_i \rangle - \langle \boldsymbol{a}_i, \boldsymbol{w} \rangle \langle \boldsymbol{a}_i, \boldsymbol{a}_k \rangle + \langle \boldsymbol{w}, \boldsymbol{a}_k \rangle \langle \boldsymbol{w}, \boldsymbol{a}_i \rangle^2 - \langle \boldsymbol{w}, \boldsymbol{a}_i \rangle \langle \boldsymbol{a}_i, \boldsymbol{a}_k \rangle \\ &= \langle \boldsymbol{w}, \boldsymbol{a}_i \rangle \left[\langle \boldsymbol{a}_i, \boldsymbol{a}_k \rangle \left(\|\boldsymbol{w}\|^2 - 2 \right) + \langle \boldsymbol{w}, \boldsymbol{a}_k \rangle \langle \boldsymbol{w}, \boldsymbol{a}_i \rangle \right]. \end{aligned}$$

From here, we see that

$$\begin{aligned}\langle \boldsymbol{w}_{t+1}, \boldsymbol{a}_k\rangle &= \langle \boldsymbol{w}_t, \boldsymbol{a}_k\rangle - \eta \sum_{i\in\mathcal{B}_t} \langle \nabla \ell(\boldsymbol{w}_t; \boldsymbol{a}_i), \boldsymbol{a}_k\rangle \\ &\overset{(i)}{=} \langle \boldsymbol{w}_t, \boldsymbol{a}_k\rangle - \eta \left[\langle \boldsymbol{w}_t, \boldsymbol{a}_k\rangle \left(\|\boldsymbol{w}_t\|^2 - 2 + \langle \boldsymbol{w}_t, \boldsymbol{a}_k\rangle^2\right) \mathbb{1}(k \in \mathcal{B}_t) + \langle \boldsymbol{w}_t, \boldsymbol{a}_k\rangle \sum_{i\in\mathcal{B}_t,\, i\neq k} \langle \boldsymbol{w}_t, a_i\rangle^2\right].\end{aligned}$$

Equality (i) uses that the $\boldsymbol{a}_\ell$'s are orthogonal. Thus, if $k \notin \mathcal{B}_t$, we see that

$$\langle \boldsymbol{w}_{t+1}, \boldsymbol{a}_k\rangle = \langle \boldsymbol{w}_t, \boldsymbol{a}_k\rangle \left(1 - \eta \sum_{i\in\mathcal{B}_t} \langle \boldsymbol{w}_t, a_i\rangle^2\right) = \langle \boldsymbol{w}_t, \boldsymbol{a}_k\rangle (1 - u_t).$$

On the other hand, if $k \in \mathcal{B}_t$, we have,

$$\begin{aligned}\langle \boldsymbol{w}_{t+1}, \boldsymbol{a}_k\rangle &= \langle \boldsymbol{w}_t, \boldsymbol{a}_k\rangle - \eta \left[\langle \boldsymbol{w}_t, \boldsymbol{a}_k\rangle \left(\|\boldsymbol{w}_t\|^2 - 2 + \langle \boldsymbol{w}_t, \boldsymbol{a}_k\rangle^2\right) + \langle \boldsymbol{w}_t, \boldsymbol{a}_k\rangle \sum_{i\in\mathcal{B}_t,\, i\neq k} \langle \boldsymbol{w}_t, a_i\rangle^2\right] \\ &= \langle \boldsymbol{w}_t, \boldsymbol{a}_k\rangle \left(1 - \eta\left(\|\boldsymbol{w}_t\|^2 + u_t - 2\right)\right).\end{aligned}$$

Putting these together, we see that the updates of $c_t(i)$ for $i \in [n]$ can be written as follows:

$$\begin{aligned}c_{t+1}(i) &= c_t(i)(1 + \eta(2 - u_t - \|\boldsymbol{w}_t\|^2)) && i \in \mathcal{B}_t, \\ c_{t+1}(j) &= c_t(j)(1 - \eta u_t) && j \in [n] \setminus \mathcal{B}_t.\end{aligned}$$

■

Throughout, we will always be operating under the following assumption on the initialization and step-size of minibatch gradient descent.

Assumption 1 *Assume that $\|\boldsymbol{w}_0\| < 1$ and $\eta \leq 1/5$.*

Under this assumption we show that the minibatch GD iterates are bounded above in norm in the following proposition which is a restatement of Proposition 5 from the main paper.

Proposition 15 (Bounded Iterates) *For all t the iterates of Eq. (3) for any mini-batch sequence $(\mathcal{B}_t)_{t\geq 0}$ satisfy $\|\boldsymbol{w}_t\|^2 \leq 1 + \eta/4$.*

Proof We prove this by induction. Clearly the statement holds at $t = 0$ by Assumption 1. Now assume the statement holds at time t. For convenience define $N_t := \|\boldsymbol{w}_t\|^2$. Since the data is orthonormal we have

$$\begin{aligned}N_{t+1} &= \sum_{i\in\mathcal{B}_t} c_{t+1}(i)^2 + \sum_{j\in[n]\setminus\mathcal{B}_t} c_{t+1}(j)^2 \\ &= \sum_{i\in\mathcal{B}_t} c_t(i)^2(1 + \eta(2 - u_t - N_t))^2 + \sum_{j\in[n]\setminus\mathcal{B}_t} c_t(j)^2(1 - \eta u_t)^2 \\ &= u_t(1 + \eta(2 - u_t - N_t))^2 + (N_t - u_t)(1 - \eta u_t)^2.\end{aligned}$$

Let us consider the following function

$$f(N,u) = u(1+\eta(2-u-N))^2 + (N-u)(1-\eta u)^2.$$

Let $N_{\max} := 1+\eta/4$. It suffices to show that

$$\max_{N,u\in[0,N_{\max}]} f(N,u) \le N_{\max}.$$

Note that if $u \ge 0$, then $f(N,u)$ is a convex quadratic in N hence

$$\begin{aligned}\max_{N,u\in[0,N_{\max}]} f(N,u) &= \max_{u\in[0,N_{\max}]} \max_{N\in[0,N_{\max}]} f(N,u)\\ &= \max_{u\in[0,N_{\max}]} \{\max(f(0,u), f(N_{\max},u))\}\\ &= \max\left(\max_{u\in[0,N_{\max}]} f(0,u), \max_{u\in[0,N_{\max}]} f(N_{\max},u)\right).\end{aligned}$$

Therefore it suffices to show that

$$\max_{u\in[0,N_{\max}]} f(0,u) \le N_{\max} \quad \text{and} \quad \max_{u\in[0,N_{\max}]} f(N_{\max},u) \le N_{\max}.$$

Plugging in $N=0$ we have

$$\max_{u\in[0,N_{\max}]} f(0,u) = \max_{u\in[0,N_{\max}]} 4\eta u(1+\eta(1-u)).$$

Since $u \mapsto 4\eta u(1+\eta(1-u))$ is increasing for $u \le (1+\eta)/2\eta$ and $N_{\max} \le (1+\eta)/2\eta$,

$$\begin{aligned}\max_{u\in[0,N_{\max}]} 4\eta u(1+\eta(1-u)) &= 4\eta N_{\max}(1+\eta(1-N_{\max}))\\ &\le N_{\max}(1-\eta^2/4) \le N_{\max}.\end{aligned}$$

We can bound the other term $\max_{u\in[0,N_{\max}]} f(N_{\max},u)$ as follows

$$\begin{aligned}\max_{u\in[0,N_{\max}]} f(N_{\max},u) &= \max_{u\in[0,N_{\max}]} u(1+\eta(2-u-N_{\max}))^2 + (N_{\max}-u)(1-\eta u)^2\\ &= \max_{u\in[0,N_{\max}]} \eta u(2-N_{\max})(2-2\eta u+\eta(2-N_{\max})) + N_{\max}(1-\eta u)^2\\ &= \max_{u\in[0,N_{\max}]} \eta^2(3N_{\max}-4)u^2 + \eta u[\eta(2-N_{\max})^2 + 4(1-N_{\max})] + N_{\max}\end{aligned}$$

Observe that the first and second terms in the last line are non-positive since

$$\begin{aligned}&3N_{\max}-4 = 3+3\eta/4-4 \le 0\\ &\eta(2-N_{\max})^2 + 4(1-N_{\max}) = \eta(1-\eta/4)^2 - \eta \le 0\end{aligned}$$

hence it follows that $\max_{u\in[0,N_{\max}]} f(N_{\max},u) \le N_{\max}$ which completes the proof. ■

The previous proposition can be used to show that the individual coordinates themselves are bounded, that they obey a "sign-stability" property, and that Ψ_t is decreasing. We show this in the following corollary which restates Corollary 6 in the main paper.

Corollary 16 *Under the conditions of Proposition 15, for all times t we have $|c_t(i)| < 1$ and $\operatorname{sign}(c_t(i)) = \operatorname{sign}(c_0(i))$ for all $i \in [n]$. Furthermore, Ψ_t is monotonically decreasing.*

Proof By Proposition 15 we have $\|\boldsymbol{w}_t\|^2 \le 1 + \eta/4$ for all t. Define $u_t := \|\Pi_{\mathcal{B}_t}(\boldsymbol{w}_t)\|^2$ as before. For any $i \in \mathcal{B}_t$, the coordinate update is

$$c_{t+1}(i) = c_t(i)(1 + \eta(2 - u_t - \|\boldsymbol{w}_t\|^2)).$$

We have that $\operatorname{sign}(c_{t+1}(i)) = \operatorname{sign}(c_t(i))$ since

$$1 + \eta(2 - u_t - \|\boldsymbol{w}_t\|^2) \ge 1 + \eta(2 - 2\|\boldsymbol{w}_t\|^2) \ge 1 - \eta^2/2 > 0,$$

therefore by induction $\operatorname{sign}(c_t(i)) = \operatorname{sign}(c_0(i))$ for all t. Furthermore,

$$|c_{t+1}(i)| = |c_t(i)|(1 + \eta(2 - u_t - \|\boldsymbol{w}_t\|^2)) \le |c_t(i)|(1 + 2\eta(1 - c_t(i)^2)).$$

Using a direct calculation (see Lemma 20 for details), this implies that if $|c_t(i)| < 1$ then $|c_{t+1}(i)| < 1$, hence by induction $|c_t(i)| < 1$ for all t. Now consider $j \in [n] \setminus \mathcal{B}_t$. The coordinate update is

$$c_{t+1}(j) = c_t(j)(1 - \eta u_t)^2.$$

Note that these coordinates are mutliplied by a quantity in $(0, 1)$ since

$$1 - \eta u_t \le 1 \text{ and } 1 - \eta u_t \ge 1 - \eta\|\boldsymbol{w}_t\|^2 \ge 1 - \eta(1 + \eta/4) > 0.$$

which easily implies the remaining claims. ■

The next two lemmas we show that the iterates can never have small norm and suggest that typically the norm show grow to one. Note that in the proofs of these lemmas we use a slightly modified definition of u_t.

Lemma 17 *If $\|\boldsymbol{w}_t\|^2 \ge 1 - \varepsilon$ for some $\varepsilon \in (0, 1)$, then $\|\boldsymbol{w}_{t+1}\|^2 \ge \|\boldsymbol{w}_t\|^2 + 4\eta\varepsilon \sum_{i\in\mathcal{B}_t} c_t(i)^2$.*

Proof Let $u_t := \sum_{i\in\mathcal{B}_t} c_t(i)^2$ and $v_t := \|\boldsymbol{w}_t\|^2 - u_t = \sum_{j\in[n]\setminus\mathcal{B}_t} c_t(j)^2$. Then we can lower bound the increments as follows

$$\begin{aligned}
\sum_{i\in\mathcal{B}_t} c_{t+1}(i)^2 - \sum_{i\in\mathcal{B}_t} c_t(i)^2 &= \sum_{i\in\mathcal{B}_t} c_t(i)^2(1 + \eta(2 - u_t - \|\boldsymbol{w}_t\|^2))^2 - c_t(i)^2 \\
&= u_t(1 + \eta(2 - 2u_t - v_t))^2 - u_t \\
&= \eta u_t(2 - 2u_t - v_t)(2 + \eta(2 - 2u_t - v_t)) \\
&= 2\eta u_t(2 - 2u_t - v_t) + \eta^2 u_t(2 - 2u_t - v_t)^2 \\
&\ge 2\eta u_t(2 - 2u_t - v_t), \\
\sum_{j\in[n]\setminus\mathcal{B}_t} c_{t+1}(j)^2 - \sum_{j\in[n]\setminus\mathcal{B}_t} c_t(j)^2 &= \sum_{j\in[n]\setminus\mathcal{B}_t} c_t(j)^2(1 - \eta u_t)^2 - c_t(j)^2 \\
&= v_t((1 - \eta u_t)^2 - 1) \\
&= -\eta v_t u_t(2 - \eta u_t) \ge -2\eta v_t u_t.
\end{aligned}$$

Combining the bounds we have

$$\begin{aligned}\|\boldsymbol{w}_{t+1}\|^2 - \|\boldsymbol{w}_t\|^2 &= \left(\sum_{i\in\mathcal{B}_t} c_{t+1}(i)^2 - \sum_{i\in\mathcal{B}_t} c_t(i)^2\right) + \left(\sum_{j\in[n]\setminus\mathcal{B}_t} c_{t+1}(j)^2 - \sum_{j\in[n]\setminus\mathcal{B}_t} c_t(j)^2\right)\\ &\geq 2\eta u_t(2-2u_t-2v_t)\\ &\geq 4\eta u_t(1-\|\boldsymbol{w}_t\|^2).\end{aligned}$$

The conclusion now follows since by assumption $1-\|\boldsymbol{w}_t\|^2 \geq \varepsilon$. ■

Lemma 18 *If $\|\boldsymbol{w}_t\| \geq 1$, then $\|\boldsymbol{w}_{t+1}\| \geq 1$.*

Proof Let $N_t := \|\boldsymbol{w}_t\|^2$ and as in the previous lemma define $u_t := \|\Pi_{\mathcal{B}_t}(\boldsymbol{w}_t)\|^2$ and $v_t := N_t - u_t$. Now observe that

$$\begin{aligned}N_{t+1} = \sum_{i\in\mathcal{B}_t} c_{t+1}(i)^2 + \sum_{j\in[n]\setminus\mathcal{B}_t} c_{t+1}(j)^2 &= u_t(1+\eta(2-2u_t-v_t))^2 + v_t(1-\eta u_t)^2\\ &\geq u_t(1+2\eta(2-2u_t-v_t)) + v_t(1-2\eta u_t)\\ &= u_t+v_t+2\eta u_t(2-2u_t-2v_t).\end{aligned}$$

The above inequality can be written as

$$N_{t+1} \geq N_t + 4\eta u_t(1-N_t).$$

The claim now follows since $N_t + 4\eta u_t(1-N_t) \geq 1$. Indeed note that

$$N_t + 4\eta u_t(1-N_t) \geq 1$$

if and only if

$$(1-4\eta u_t)(N_t-1) \geq 0$$

which is true since $N_t \geq 1$ and by Proposition 15

$$4\eta u_t \leq 4\eta N_t \leq 4\eta(1+\eta/4) \leq (4/5)\cdot(1+1/20) \leq 1$$

since $\eta \leq 1/5$ by Assumption 1. ■

Lastly, we show that $\Phi_t = \|\Pi_m(\boldsymbol{w}_t)\|^2$ is always lower bounded by a constant.

Lemma 19 *We have $\Phi_t \geq \delta$ for all t, where we define the constant*

$$\delta := \min(\Phi_0, 1-\Psi_0) > 0. \tag{20}$$

Proof While $\|\boldsymbol{w}_t\|^2 = \Phi_t + \Psi_t < 1$ we must have Φ_t increasing since $\|\boldsymbol{w}_t\|^2$ is increasing by Lemma 17 and Ψ_t is decreasing by Corollary 16, hence $\Phi_t \geq \Phi_0$. If at some point $\|\boldsymbol{w}_t\|^2 \geq 1$, then by Lemma 18 for all t thereafter $\Phi_t \geq 1-\Psi_t \geq 1-\Psi_0$. Combining these two cases gives that $\Phi_t \geq \delta$ as desired. ■

A.1 Technical Lemmas

Lemma 20 *Let* $f(x) = x(1+\lambda(1-x^2))$. *If* $\lambda \in (0, 1/2]$, *then* $f(x) \in (0,1)$ *for all* $x \in (0,1)$.

Proof Computing the derivative $f'(x) = 1 + \lambda - 3\lambda x^2$. Note that $f'(x) > 0$ iff

$$x^2 < \frac{1+\lambda}{3\lambda}.$$

For $0 < \lambda \leq 1/2$ we have

$$\frac{1+\lambda}{3\lambda} \geq 1,$$

therefore $f'(x) > 0$ for $x \in (0,1)$. Thus for $x \in (0,1)$, $0 = f(0) < f(x) < f(1) = 1$. ■

Appendix B. General Properties of Minibatch SGD

Now we will present some general properties of minibatch SGD on linear autoencoders. By minibatch SGD, we are referring to the minibatch GD algorithm where at each iteration t, the minibatch $\mathcal{B}_t$ chosen uniformly from the subsets of $[m]$ of size b. In contrast, the results in the previous section hold for any sequence of minibatch indices.

Here and in the next section the maximal index

$$i_t^\star := \arg\max_{i\in[m]} |c_t(i)| \tag{21}$$

will play an important role in the analysis of SGD. The first result for this section shows that $\psi_t \to 0$.

Proposition 21 *As $t \to \infty$, almost surely $\Psi_t \to 0$.*

Proof Define $u_t := \|\Pi_{\mathcal{B}_t}(\boldsymbol{w}_t)\|^2$. By Corollary 16, Ψ_t is decreasing and hence converges to some limiting value $\Psi^\star$. For the sake of contradiction, assume that $\Psi^\star > 0$. Then almost surely $i_t^\star \in \mathcal{B}_t$ infinitely often. Therefore by Lemma 19, $u_t \geq c_t(i_t^\star)^2 \geq \Phi_t/m \geq \delta/m$, infinitely often with δ defined in Eq. (20). Let $T(\varepsilon)$ be a time such that

$$\Psi_T \leq (1+\varepsilon)\Psi^\star.$$

Then almost surely, there exists some $t \geq T$ such that $u_t \geq \Phi_t/m \geq \delta/m$, hence

$$\Psi_{t+1} = \Psi_t(1-\eta u_t)^2 \leq \Psi_T(1-\eta\delta/m)^2 \leq (1+\varepsilon)(1-\eta\delta/m)\Psi^\star.$$

If we take $\varepsilon = \eta\delta/m$, then $\Psi_{t+1} < \Psi^\star$ which is a contradiction hence $\Psi^\star = 0$. ■

Proposition 22 *Almost surely $\liminf_{t\to\infty} \|\boldsymbol{w}_t\| \geq 1$.*

Proof If $\|\boldsymbol{w}_t\| \geq 1$ at some time t, then the claim immediately follows from Lemma 18. Therefore let us assume that $\|\boldsymbol{w}_t\| < 1$ for all t. For the sake of contradiction assume that $\liminf\limits_{t\to\infty} \|\boldsymbol{w}_t\| < 1$. Therefore there exists some $\tilde{\varepsilon} > 0$ such that $\|\boldsymbol{w}_t\| \geq 1-\tilde{\varepsilon}$ at most finitely many times. Since we assumed $\|\boldsymbol{w}_t\| < 1$ for all t, in fact there exists $\varepsilon > 0$ such that $\|\boldsymbol{w}_t\|^2 \leq 1-\varepsilon$ for all t. Almost surely there exists a countably infinite set of times $(t_k)_{k=0}^\infty$ such that $i_{t_k} = i_{t_k}^\star$. By Lemma 17 we have that $\|\boldsymbol{w}_t\|$ is monotonically increasing for all t and that for any index $s = t_k$,

$$\|\boldsymbol{w}_{s+1}\|^2 - \|\boldsymbol{w}_s\|^2 \geq 4\eta\varepsilon \sum_{i\in\mathcal{B}_s} c_t(i)^2 \geq 4\eta\varepsilon c_t(i_s^\star)^2 \geq 4\eta\varepsilon\Phi_t/m \geq 4\eta\varepsilon\delta/m,$$

where the last inequality is by Lemma 19 with δ defined in Eq. (20). Hence it is clear that $\|\boldsymbol{w}_t\| \to \infty$ which is a contradiction and so $\liminf\limits_{t\to\infty} \|\boldsymbol{w}_t\| \geq 1$. ■

Appendix C. Minibatch SGD Convergence

In this section we will prove the convergence theorem for minibatch SGD stated in Theorem 2. Define $\mathbb{Z}_+ := \{0, 1, \ldots\}$ to be the set of nonnegative integers and $\mathbb{R}_+ := [0, \infty)$ to be the set of nonnegative reals. Recall the random variable R_t which is defined as

$$R_t := \log \left(\frac{|c_t(i_t^\star)|}{\sum\limits_{\ell \in [m] \setminus \{i_t^\star\}} |c_t(\ell)|} \right), \tag{22}$$

where $i_t^\star$ is defined in Eq. (21). A key step will be to prove Proposition 9 which establishes the transience of the stochastic process $R = (R_t)_{t \in \mathbb{Z}^+}$. From there, we will be able to give the convergence behavior of the iterates $\boldsymbol{w}_t$ by invoking results from previous sections.

C.1 Transience of R

For the purpose of analysing the stochastic process R we can slightly simplify matters and for the proof of Proposition 9 given in this section assume the following without loss of generality,

Assumption 2 *$c_t(\ell) > 0$ for all $\ell \in [n]$ and for all t.*

To see why, observe that for any initialization $(c_0(1), \ldots, c_0(n))$, we can also consider a coupled "parallel" trajectory induced by running minibatch SGD on the initialization $(\widetilde{c}_0(1), \ldots, \widetilde{c}_0(n))$ where $\widetilde{c}_0(\ell) = c_0(\ell) \cdot \mathrm{sign}(c_0(\ell))$ using the same minibatch sequence. By Corollary 16 it is not hard to see that $|c_t(\ell)| = c_t(\ell) \cdot \mathrm{sign}(c_0(\ell)) = \widetilde{c}_t(\ell)$ for all t, hence $R_t = \widetilde{R}_t$ where $\widetilde{R}_t$ is the stochastic process in Eq. (22) induced by the parallel trajectory. Note that $\widetilde{c}_t(\ell)$ satisfies Assumptions 1 and 2, and that R_t is transient if and only if $\widetilde{R}_t$ is transient, therefore it suffices to analyse trajectories obeying Assumption 2.

To establish the transience of the stochastic process R we use a general result from non-homogeneous random walk theory stated in Proposition 8 which gives general conditions for transience. Recall that $\mathcal{B}_t \subseteq [m]$ is the selected minibatch indices at time t where where each minibatch is selected at uniform from $\{\mathcal{B} \subseteq [m] : |\mathcal{B}| = b\}$. Let $\mathcal{F}_t = \sigma(\mathcal{B}_0, \ldots, \mathcal{B}_t)$ be the sigma-algebra generated by the random minibatch draws $\mathcal{B}_0, \ldots, \mathcal{B}_t$. Clearly the stochastic process R is adapted to the filtration $(\mathcal{F}_t)_{t \in \mathbb{Z}^+}$.

Note that $c_t(i_t^\star) \geq c_t(\ell)$ for any $\ell \in [m]$ by definition of $i_t^\star$ and so we have the deterministic lower bound $R_t \geq -\log(m)$. Therefore R is a stochastic process on the (translated) half-line. To invoke the theorem directly we will have to verify that R obeys assumptions (A1) and (A2) as well as lower bound the conditional mean increment, that is, show that there exists a function $\underline{\mu}_1 : \mathbb{R} \to \mathbb{R}$ such that for all $t \in \mathbb{Z}^+$

$$\underline{\mu}_1(R_t) \leq \mathbb{E}(R_{t+1} - R_t \mid \mathcal{F}_t), \quad \text{a.s.},$$

and $\liminf\limits_{r \to \infty} \underline{\mu}_1(r) > 0$. For ease of presentation we lower bound the conditional increment first and then verify the assumptions. To this end we will show the following.

Proposition 23 *For all $t \in \mathbb{Z}_+$, there exists $\delta_1 : \mathbb{R} \to \mathbb{R}_+$ and $\delta_2 : \mathbb{R}^n \to \mathbb{R}_+$ such that*

$$\mathbb{E}(R_{t+1} - R_t \mid \mathcal{F}_t) \geq \frac{\eta b (m-b)^2}{2m(m-1)^2} - \delta_1(R_t) - \delta_2(\boldsymbol{w}_t) - \delta_1(R_t)\delta_2(\boldsymbol{w}_t), \quad \textit{a.s.},$$

$\delta_1(r) \to 0$ *as* $r \to \infty$*, and* $\delta_2(\boldsymbol{w}_t) \to 0$ *almost surely as* $t \to \infty$*.*

Note that this is actually sufficient for our purposes since we can apply the theorem to the tail process $(R_t)_{t \geq \tau(\varepsilon)}$ where $\tau(\varepsilon)$ defined as the stopping time

$$\tau(\varepsilon) := \inf\{\tau \in \mathbb{Z}^+ : \delta_2(\boldsymbol{w}_t) \leq \varepsilon, \text{ for all } t \geq \tau\}, \quad \varepsilon = \frac{\eta b(m-b)^2}{4m(m-1)^2},$$

since the proposition gives that for all $t \geq \tau(\varepsilon)$

$$\underline{\mu}_1(R_t) := \frac{\eta b(m-b)^2}{2m(m-1)^2} - \delta_1(R_t) - \varepsilon - \varepsilon\delta_1(R_t) \leq \mathbb{E}(R_{t+1} - R_t \mid \mathcal{F}_t), \quad \text{a.s.},$$

and that

$$\liminf_{t\to\infty} \underline{\mu}_1(r) = \liminf_{t\to\infty} \frac{\eta b(m-b)^2}{2m(m-1)^2} - \delta_1(r) - \varepsilon - \varepsilon\delta_1(r) = \frac{\eta b(m-b)^2}{2m(m-1)^2} - \varepsilon = \varepsilon > 0.$$

To obtain the lower bound in Proposition 23 we will bound the increment at time t

$$\Delta_t := R_{t+1} - R_t = (R_{t+1} - R_t)\mathbb{1}(i_t^\star \in \mathcal{B}_t) + (R_{t+1} - R_t)\mathbb{1}(i_t^\star \notin \mathcal{B}_t),$$

by considering two cases.

1. In the first case, $i_t^\star \in \mathcal{B}_t$ and we will give a positive lower bound on Δ_t.
2. In the second case, $i_t^\star \notin \mathcal{B}_t$ and we will upper bound how negative Δ_t can be.

By averaging over these two cases we will obtain an asymptotically positive lower bound $\underline{\mu}_1$ as desired. Now let us introduce some notation that will be useful in this section. Define

$$\mathcal{J}_t := [m] \setminus \{i_t^\star\}, \qquad S_t := \sum_{\ell \in \mathcal{J}_t} c_t(\ell)$$

$$u_t(\mathcal{B}) := \sum_{\ell \in \mathcal{B}} c_t(\ell)^2, \qquad \Psi_t := \sum_{\ell \in [n]\setminus[m]} c_t(\ell)^2$$

Note that $S_t = c_t(i_t^\star) \cdot \exp(-R_t)$ by definition of R_t. We also define the update factors

$$A_t(\mathcal{B}) = 1 + \eta(2 - u_t(\mathcal{B}) - \|\boldsymbol{w}_t\|^2), \quad B_t(\mathcal{B}) = 1 - \eta u_t(\mathcal{B}).$$

Observe that $A_t(\mathcal{B})$ and $B_t(\mathcal{B})$ are precisely the multiplicative factors such that

$$\begin{aligned} c_{t+1}(i) &= c_t(i) \cdot A_t(\mathcal{B}_t) && \text{if } i \in \mathcal{B}_t, \\ c_{t+1}(j) &= c_t(j) \cdot B_t(\mathcal{B}_t) && \text{if } j \notin \mathcal{B}_t. \end{aligned}$$

Denote $\overline{\mathcal{B}} = [m] \setminus \mathcal{B}$. Then we define the final set of quantities

$$X_t(\mathcal{B}) = \sum_{\ell \in \mathcal{B}} c_t(\ell), \qquad Y_t(\mathcal{B}) = \sum_{\ell \in \overline{\mathcal{B}}} c_t(\ell),$$

$$\widetilde{X}_t(\mathcal{B}) = \sum_{\ell \in \mathcal{B}\setminus\{i_t^\star\}} c_t(\ell), \qquad \widetilde{Y}_t(\mathcal{B}) = \sum_{\ell \in \overline{\mathcal{B}}\setminus\{i_t^\star\}} c_t(\ell).$$

Clearly, if $i_t^\star \in \mathcal{B}$ then $S_t = \widetilde{X}_t(\mathcal{B}) + Y_t(\mathcal{B})$ and $S_t = X_t(\mathcal{B}) + \widetilde{Y}_t(\mathcal{B})$ otherwise. For convenience we use $o(R_t)$ and $o(t)$ to make the following substitutions

$$\begin{aligned} o(R_t) &\equiv f(R_t), \quad f : \mathbb{R} \to \mathbb{R}_+, \quad \lim_{r\to\infty} f(r) = 0 \\ o(t) &\equiv g(\boldsymbol{w}_t), \quad g : \mathbb{R}^n \to \mathbb{R}_+, \quad \lim_{t\to\infty} g(\boldsymbol{w}_t) = 0, \text{ a.s.} \end{aligned}$$

For example, we will write $o(R_t)$ in place of $\exp(-R_t)$. We will also use

$$\mathfrak{o}(t, R_t) \equiv o(R_t) + o(t) + o(t)o(R_t)$$

to make further substitutions when terms of this form which arise. Before giving the proof of Proposition 23 we give a lemma which gives asymptotic bounds on the ratio of $A_t(\mathcal{B})/B_t(\mathcal{B})$ which will important later for lower bounding the conditional mean increment.

Lemma 24 *Let $\mathcal{B} \subseteq [m]$ be a set of minibatch indices. If $i_t^\star \in \mathcal{B}$ then*

$$\frac{A_t(\mathcal{B})}{B_t(\mathcal{B})} \geq \frac{1}{1-\eta} - o(R_t) - o(t) = \frac{1}{1-\eta} - \mathfrak{o}(t, R_t),$$

and if $i_t^\star \notin \mathcal{B}$ then

$$\frac{A_t(\mathcal{B})}{B_t(\mathcal{B})} \leq 1 + \eta + o(t) + o(R_t) + o(t)o(R_t) = 1 + \eta + \mathfrak{o}(t, R_t).$$

Proof To begin with it will be helpful to recall that $c_t(\ell) < 1$ for all $\ell \in [n]$ by Corollary 16 and $\Psi_t = o(t)$ by Proposition 21, therefore

$$S_t = \sum_{\ell\in\mathcal{J}_t} c_t(\ell) = c_t(i_t^\star)\exp(-R_t) \leq \exp(-R_t) = o(R_t)$$

and furthermore

$$\begin{aligned} \|\boldsymbol{w}_t\|^2 &= c_t(i_t^\star)^2 + \sum_{\ell\in\mathcal{J}_t} c_t(\ell)^2 + \Psi_t \\ &\leq c_t(i_t^\star)^2 + \sum_{\ell\in\mathcal{J}_t} c_t(\ell) + \Psi_t \\ &\leq 1 + o(R_t) + o(t). \end{aligned}$$

Now let us start with the first inequality we we wish to show. Assume that $i_t^\star \in \mathcal{B}$. Then,

$$\begin{aligned} \frac{A_t(\mathcal{B})}{B_t(\mathcal{B})} &= \frac{1 + \eta(2 - u_t(\mathcal{B}) - \|\boldsymbol{w}_t\|^2)}{1 - \eta u_t(\mathcal{B})} \\ &= \frac{1 + \eta(2 - u_t(\mathcal{B}))}{1 - \eta u_t(\mathcal{B})} - \frac{\eta}{1 - \eta u_t(\mathcal{B})}\|\boldsymbol{w}_t\|^2 \\ &= \frac{1 + \eta(2 - u_t(\mathcal{B}))}{1 - \eta u_t(\mathcal{B})} - \frac{\eta}{1 - \eta u_t(\mathcal{B})}(u_t(\mathcal{B}) + \|\boldsymbol{w}_t\|^2 - u_t(\mathcal{B})) \\ &= \frac{1 + 2\eta(1 - u_t(\mathcal{B}))}{1 - \eta u_t(\mathcal{B})} - \frac{\eta}{1 - \eta u_t(\mathcal{B})}(\|\boldsymbol{w}_t\|^2 - u_t(\mathcal{B})) \\ &\geq \frac{1 + 2\eta(1 - u_t(\mathcal{B}))}{1 - \eta u_t(\mathcal{B})} - \frac{\eta}{1 - 2\eta}(\|\boldsymbol{w}_t\|^2 - u_t(\mathcal{B})) \end{aligned}$$

where the last line uses the fact that $u_t(\mathcal{B}) \leq \|\boldsymbol{w}_t\|^2 \leq 2$ by Proposition 15. Since $i_t^\star \in \mathcal{B}$,

$$\|\boldsymbol{w}_t\|^2 - u_t(\mathcal{B}) \leq S_t + \Psi_t = o(R_t) + o(t).$$

Combing with the above we have

$$\frac{A_t(\mathcal{B})}{B_t(\mathcal{B})} \geq \frac{1 + 2\eta(1 - u_t(\mathcal{B}))}{1 - \eta u_t(\mathcal{B})} - o(R_t) - o(t).$$

Consider the function

$$f(x) = \frac{1 + 2\eta(1 - x)}{1 - \eta x}.$$

A simple computation yields that if $\eta < 1/2$ then $f(x)$ is decreasing for all $x \in \mathbb{R}$ since

$$f'(x) = \frac{\eta(2\eta - 1)}{(1 - \eta x)^2} < 0.$$

Therefore since $u_t(\mathcal{B}) \leq \|\boldsymbol{w}_t\|^2 \leq 1 + o(R_t) + o(t)$

$$\begin{aligned}\frac{1 + 2\eta(1 - u_t(\mathcal{B}))}{1 - \eta u_t(\mathcal{B})} &\geq f(1 + o(R_t) + o(t)) \\ &\geq f(1) - o(R_t) - o(t) \\ &= \frac{1}{1 - \eta} - o(R_t) - o(t).\end{aligned}$$

Therefore combining everything together yields the first desired inequality

$$\frac{A_t(\mathcal{B})}{B_t(\mathcal{B})} \geq \frac{1}{1 - \eta} - o(R_t) - o(t).$$

For the other inequality, assume now that $i_t^\star \notin \mathcal{B}$. Then,

$$\begin{aligned}\frac{A_t(\mathcal{B})}{B_t(\mathcal{B})} &= \frac{1 + \eta(2 - u_t(\mathcal{B}) - \|\boldsymbol{w}_t\|^2)}{1 - \eta u_t(\mathcal{B})} \\ &= 1 + \eta\frac{2 - \|\boldsymbol{w}_t\|^2}{1 - \eta u_t(\mathcal{B})} \\ &\leq 1 + \eta\frac{2 - \|\boldsymbol{w}_t\|^2}{1 - o(R_t)}\end{aligned}$$

where the last line follows from the fact that $u_t(\mathcal{B}) \leq S_t = o(R_t)$. Now observe that by Proposition 22 we have $\|\boldsymbol{w}_t\|^2 \geq 1 - o(t)$, therefore

$$\begin{aligned}\frac{A_t(\mathcal{B})}{B_t(\mathcal{B})} &\leq 1 + \eta\frac{1 + o(t)}{1 - o(R_t)} \\ &\leq 1 + \eta(1 + o(t))(1 + o(R_t)) \\ &= 1 + \eta + o(t) + o(R_t) + o(t)o(R_t).\end{aligned}$$

■

We are now ready to give the proof of Proposition 23.
Proof [Proposition 23] First observe that

$$\sum_{\ell\in\mathcal{J}_{t+1}} c_{t+1}(\ell) = \sum_{\ell\in[m]} c_{t+1}(\ell) - c_{t+1}(i^\star_{t+1}) \leq \sum_{\ell\in[m]} c_{t+1}(\ell) - c_{t+1}(i^\star_t) = \sum_{\ell\in\mathcal{J}_t} c_{t+1}(\ell)$$

since $c_{t+1}(i^\star_{t+1}) \geq c_{t+1}(i^\star_t)$, hence we have that

$$\begin{aligned} R_{t+1} - R_t &= \log\left(\frac{c_{t+1}(i^\star_{t+1})}{\sum\limits_{\ell\in\mathcal{J}_{t+1}} c_{t+1}(\ell)}\right) - \log\left(\frac{c_t(i^\star_t)}{\sum\limits_{\ell\in\mathcal{J}_t} c_t(\ell)}\right) \\ &\geq \log\left(\frac{c_{t+1}(i^\star_t)}{\sum\limits_{\ell\in\mathcal{J}_t} c_{t+1}(\ell)}\right) - \log\left(\frac{c_t(i^\star_t)}{\sum\limits_{\ell\in\mathcal{J}_t} c_t(\ell)}\right). \end{aligned} \tag{23}$$

We start with the case $i^\star_t \in \mathcal{B}_t$. Starting from Eq. (23), we have that on the event $i^\star_t \in \mathcal{B}_t$,

$$\begin{aligned} R_{t+1} - R_t &\geq \log\left(\frac{c_{t+1}(i^\star_t)}{\sum\limits_{\ell\in\mathcal{J}_t} c_{t+1}(\ell)}\right) - \log\left(\frac{c_t(i^\star_t)}{\sum\limits_{\ell\in\mathcal{J}_t} c_t(\ell)}\right) \\ &= \log\left(\frac{c_t(i^\star_t)A_t(\mathcal{B}_t)}{A_t(\mathcal{B}_t)\widetilde{X}_t(\mathcal{B}_t) + B_t(\mathcal{B}_t)Y_t(\mathcal{B}_t)}\right) - \log\left(\frac{c_t(i^\star_t)}{S_t}\right) \\ &= \log\left(\frac{A_t(\mathcal{B}_t)S_t}{A_t(\mathcal{B}_t)(S_t - Y_t(\mathcal{B}_t)) + B_t(\mathcal{B}_t)Y_t(\mathcal{B}_t)}\right) \\ &= \log\left(\frac{S_t}{S_t - (1 - B_t(\mathcal{B}_t)/A_t(\mathcal{B}_t))\cdot Y_t(\mathcal{B}_t)}\right). \end{aligned} \tag{24}$$

By Lemma 24,

$$\frac{B_t(\mathcal{B}_t)}{A_t(\mathcal{B}_t)} \leq 1 - \eta + \mathfrak{o}(t, R_t),$$

hence the above yields

$$(R_{t+1} - R_t)\cdot \mathbb{1}(i^\star_t \in \mathcal{B}_t) \geq \log\left(\frac{S_t}{S_t - [\eta - \mathfrak{o}(t,R_t)]\cdot Y_t(\mathcal{B})}\right)\cdot \mathbb{1}(i^\star_t \in \mathcal{B}_t).$$

Now we can apply Jensen's inequality and obtain that

$$\mathbb{E}(R_{t+1} - R_t \mid i^\star_t \in \mathcal{B}_t, \mathcal{F}_{t-1}) \geq \log\left(\frac{S_t}{S_t - [\eta - \mathfrak{o}(t,R_t)]\cdot \mathbb{E}(Y_t(\mathcal{B}) \mid i^\star_t \in \mathcal{B}_t, \mathcal{F}_{t-1})}\right),$$

Defining $\overline{\mathcal{B}}_t := [m]\setminus\mathcal{B}_t$, we can compute

$$\begin{aligned} \mathbb{E}(Y_t(\mathcal{B}) \mid i^\star_t \in \mathcal{B}_t, \mathcal{F}_{t-1}) &= \mathbb{E}\left(\sum_{\ell\in[m]} \mathbb{1}(\ell\in\overline{\mathcal{B}}_t)\cdot c_t(\ell) \;\middle|\; i^\star_t \in \mathcal{B}_t, \mathcal{F}_{t-1}\right) \\ &= \sum_{\ell\in[m]} \Pr\bigl(\ell\in\overline{\mathcal{B}}_t \mid i^\star_t\in\mathcal{B}_t\bigr)\cdot c_t(\ell) \end{aligned}$$

Recall that $\mathcal{B}_t$ is uniformly chosen from the set of subsets of $[m]$ of size b. Therefore conditional on $i_t^\star \in \mathcal{B}_t$, $\overline{\mathcal{B}}_t$ is uniformly chosen from the set of subsets of $[m] \setminus \{i_t^\star\}$ of size $m-b$, hence $\Pr\bigl(\ell \in \overline{\mathcal{B}}_t \mid i_t^\star \in \mathcal{B}_t\bigr)$ is $(m-b)/(m-1)$ if $\ell \neq i_t^\star$ and 0 otherwise. Therefore

$$\begin{aligned}
\mathbb{E}(Y_t(\mathcal{B}) \mid i_t^\star \in \mathcal{B}_t, \mathcal{F}_{t-1}) &= \sum_{\ell \in [m]} \Pr\bigl(\ell \in \overline{\mathcal{B}}_t \mid i_t^\star \in \mathcal{B}_t\bigr) \cdot c_t(\ell) \\
&= \frac{m-b}{m-1} \sum_{\ell \in [m]\setminus\{i_t^\star\}} c_t(\ell) = \frac{m-b}{m-1} S_t,
\end{aligned}$$

which then leads to

$$\begin{aligned}
\mathbb{E}(R_{t+1} - R_t \mid i_t^\star \in \mathcal{B}_t, \mathcal{F}_{t-1}) &\geq \log\left(\frac{S_t}{S_t - [\eta - \mathfrak{o}(t, R_t)] \cdot \frac{m-b}{m-1} \cdot S_t}\right) \\
&= \log\left(\frac{1}{1 - \eta \cdot \frac{m-b}{m-1}}\right) - \mathfrak{o}(t, R_t).
\end{aligned}$$

Now consider the case where $i_t^\star \notin \mathcal{B}$. Again using Eq. (23) we have that on this event,

$$\begin{aligned}
-(R_{t+1} - R_t) &\leq -\log\left(\frac{c_{t+1}(i_t^\star)}{\sum\limits_{\ell \in \mathcal{J}_t} c_{t+1}(\ell)}\right) + \log\left(\frac{c_t(i_t^\star)}{\sum\limits_{\ell \in \mathcal{J}_t} c_t(\ell)}\right) \\
&= \log\left(\frac{\widetilde{Y}_t(\mathcal{B}_t) B_t(\mathcal{B}_t) + X_t(\mathcal{B}_t) A_t(\mathcal{B}_t)}{c_t(i_t^\star) B_t(\mathcal{B}_t)}\right) + \log\left(\frac{c_t(i_t^\star)}{S_t}\right) \\
&= \log\left(\widetilde{Y}_t(\mathcal{B}_t) + X_t(\mathcal{B}_t) \frac{A_t(\mathcal{B}_t)}{B_t(\mathcal{B}_t)}\right) - \log(S_t).
\end{aligned}$$

By Lemma 24 we have that if $i_t^\star \notin \mathcal{B}_t$ then

$$\frac{A_t(\mathcal{B}_t)}{B_t(\mathcal{B}_t)} \leq 1 + \eta + \mathfrak{o}(t, R_t).$$

Hence we can bound the above as

$$\begin{aligned}
-(R_{t+1} - R_t) &= \log\left(\widetilde{Y}_t(\mathcal{B}_t) + X_t(\mathcal{B}_t) \cdot [1 + \eta + \mathfrak{o}(t, R_t)]\right) - \log(S_t) \\
&= \log\left(\widetilde{Y}_t + X_t(\mathcal{B}_t) + X_t(\mathcal{B}_t) \cdot [\eta + \mathfrak{o}(t, R_t)]\right) - \log(S_t) \\
&= \log\left(S_t + X_t(\mathcal{B}_t) \cdot [\eta + \mathfrak{o}(t, R_t)]\right) - \log(S_t) \\
&= \log\left(1 + \frac{X_t(\mathcal{B}_t)}{S_t} \cdot [\eta + \mathfrak{o}(t, R_t)]\right) \\
&\leq \frac{X_t(\mathcal{B}_t)}{S_t} \cdot [\eta + \mathfrak{o}(t, R_t)].
\end{aligned}$$

Thus we have shown that

$$-(R_{t+1} - R_t) \cdot \mathbb{1}(i_t^\star \notin \mathcal{B}_t) \leq \frac{X_t(\mathcal{B}_t)}{S_t} \cdot [\eta + \mathfrak{o}(t, R_t)] \cdot \mathbb{1}(i_t^\star \notin \mathcal{B}_t).$$

Therefore by negating and taking expectations on both sides

$$\mathbb{E}(R_{t+1} - R_t \mid i_t^\star \notin \mathcal{B}, \mathcal{F}_{t-1}) \geq -(\eta + \mathfrak{o}(t, R_t)) \cdot \frac{\mathbb{E}(X_t(\mathcal{B}) \mid i_t^\star \notin \mathcal{B}_t, \mathcal{F}_{t-1})}{S_t}.$$

We can compute the conditional expectation

$$\begin{aligned}
\mathbb{E}(X_t(\mathcal{B}) \mid i_t^\star \notin \mathcal{B}_t, \mathcal{F}_{t-1}) &= \mathbb{E}\left(\sum_{\ell \in [m]} \mathbb{1}(\ell \in \mathcal{B}_t) \cdot c_t(\ell) \;\middle|\; i_t^\star \notin \mathcal{B}_t, \mathcal{F}_{t-1} \right) \\
&= \sum_{\ell \in [m]} \Pr(\ell \in \mathcal{B}_t \mid i_t^\star \notin \mathcal{B}_t) \cdot c_t(\ell) \\
&= \frac{b}{m-1} \sum_{\ell \in [m] \setminus \{i_t^\star\}} c_t(\ell) = \frac{b}{m-1} S_t,
\end{aligned}$$

where we used the fact that conditional on $i_t^\star \notin \mathcal{B}_t$, $\mathcal{B}_t$ is uniformly chosen from the set of subsets of $[m] \setminus \{i_t\}$ of size b. This then gives

$$\mathbb{E}(R_{t+1} - R_t \mid i_t^\star \notin \mathcal{B}, \mathcal{F}_{t-1}) \geq -(\eta + \mathfrak{o}(t, R_t)) \frac{b}{m-1} \frac{S_t}{S_t} = -\eta \frac{b}{m-1} - \mathfrak{o}(t, R_t).$$

Therefore combining the two cases using the law of total probability gives

$$\begin{aligned}
\mathbb{E}(R_{t+1} - R_t \mid \mathcal{F}_{t-1}) &= \Pr(i_t^\star \in \mathcal{B}_t) \cdot \mathbb{E}(R_{t+1} - R_t \mid i_t^\star \in \mathcal{B}_t, \mathcal{F}_{t-1}) \\
&\quad + \Pr(i_t^\star \notin \mathcal{B}_t) \cdot \mathbb{E}(R_{t+1} - R_t \mid i_t^\star \notin \mathcal{B}_t, \mathcal{F}_{t-1}) \\
&\geq \frac{b}{m} \log\left(\frac{1}{1 - \eta \cdot \frac{m-b}{m-1}} \right) - \eta \frac{m-b}{m} \frac{b}{m-1} - \mathfrak{o}(t, R_t) \\
&\geq \eta \frac{b}{m} \frac{m-b}{m-1} \left(1 + \frac{\eta}{2} \frac{m-b}{m-1} \right) - \eta \frac{m-b}{m} \frac{b}{m-1} - \mathfrak{o}(t, R_t) \\
&= \frac{\eta b (m-b)^2}{2m(m-1)^2} - \mathfrak{o}(t, R_t),
\end{aligned}$$

where the second inequality used that

$$\log\left(\frac{1}{1-x} \right) \geq x \cdot (1 + x/2), \quad \text{for all } x \in (0,1).$$

■

Now it remains to verify Assumptions (A1) and (A2). Let us first consider (A1). Essentially (A1) will hold because, as suggested in the proof of the previous Proposition 23, R_t increases if $i_t^\star \in \mathcal{B}_t$. We will show that in this case R_t will increase by at least a constant amount for any time t. Hence R_t can grow to an arbitrarily large value with constant probability. Here constant probability means that the probability is independent of the past, but potentially dependent on the desired target value, which is what is required by the condition in (A1). We now show this holds formally.

Proof (A1) holds We will uses parts of the proof of Proposition 23 and use the same notation. If $i_t^\star \in \mathcal{B}_t$ then from Eq. (24)

$$R_{t+1} - R_t \geq \log\left(\frac{S_t}{S_t - (1 - B_t(\mathcal{B}_t)/A_t(\mathcal{B}_t)) \cdot Y_t(\mathcal{B}_t)}\right).$$

Using the fact that $Y_t(\mathcal{B}_t) \leq S_t$ we then have

$$\Delta_t(\mathcal{B}) \geq \log\left(\frac{S_t}{S_t - (1 - B_t(\mathcal{B}_t)/A_t(\mathcal{B}_t)) \cdot S_t}\right) = \log\left(\frac{\mathcal{A}_t(\mathcal{B}_t)}{B_t(\mathcal{B}_t)}\right).$$

In the proof of the proposition we used an asymptotic lower bound on $A_t(\mathcal{B}_t)/B_t(\mathcal{B}_t)$ from Lemma 24, however here we will use the simpler non-asymptotic bound

$$\begin{aligned}\frac{A_t(\mathcal{B}_t)}{B_t(\mathcal{B}_t)} &= \frac{1 + \eta(2 - u_t(\mathcal{B}_t) - \|\boldsymbol{w}_t\|^2)}{1 - \eta u_t(\mathcal{B}_t)} \\ &\geq 1 + \eta\frac{2 - \|\boldsymbol{w}_t\|^2}{1 - \eta u_t(\mathcal{B}_t)} \\ &\geq 1 + \eta(1 - \eta/4) \geq 1 + \frac{19}{20}\eta,\end{aligned}$$

which just follows since $0 \leq u_t(\mathcal{B}_t) \leq \|\boldsymbol{w}_t\|^2 \leq 1 + \eta/4$ by Proposition 15 and $\eta \leq 1/5$ by Assumption 1. Thus we have the lower bound

$$R_{t+1} - R_t \geq \log\left(1 + \frac{19}{20}\eta\right).$$

Hence we have shown that if $i_t^\star \in \mathcal{B}_t$ then $R_{t+1} - R_t \geq \delta$ for $\delta := \log(1 + (19/20)\eta)$. Now we can easily show that (A1) holds. For any $T \in \mathbb{Z}^+$ consider a sequence of minibatches $(\mathcal{B}_s)_{s=0}^T$. For any $y \in (0, \infty)$ we can take $v(T) = \lceil \delta^{-1}(y + \log(m)) \rceil$ so that if $i_T^\star \in \mathcal{B}_t$ for $T \leq t \leq v(T)$ then $R_{T+v(T)} \geq y$ since

$$R_{T+v(T)} \geq R_T + \delta v(T) \geq -\log(m) + \delta \cdot [\delta^{-1}(y + \log(m))] = y.$$

Furthermore this occurs with probability $(b/m)^{\lceil \delta^{-1}(y+\log(m)) \rceil} > 0$.

Proof (A2) holds We now show that the process $(R_t)_{t \in \mathbb{Z}^+}$ has bounded increments. For convenience define $\mathcal{J}_t := [m] \setminus \{i_t^\star\}$. By definition

$$\begin{aligned}R_{t+1} - R_t &= \log\left(\frac{c_{t+1}(i_{t+1}^\star)}{\sum\limits_{\ell \in \mathcal{J}_{t+1}} c_{t+1}(\ell)}\right) - \log\left(\frac{c_t(i_t^\star)}{\sum\limits_{\ell \in \mathcal{J}_t} c_t(\ell)}\right) \\ &= \log\left(\frac{c_{t+1}(i_{t+1}^\star)}{c_t(i_t^\star)} \frac{\sum\limits_{\ell \in \mathcal{J}_t} c_t(\ell)}{\sum\limits_{\ell \in \mathcal{J}_{t+1}} c_{t+1}(\ell)}\right).\end{aligned}$$

It then suffices to show that the quantity I defined as

$$I := I_1 \cdot I_2, \quad I_1 = \frac{c_{t+1}(i^\star_{t+1})}{c_t(i^\star_t)}, I_2 = \frac{\sum\limits_{\ell\in\mathcal{J}_t} c_t(\ell)}{\sum\limits_{\ell\in\mathcal{J}_{t+1}} c_{t+1}(\ell)},$$

lies in a time-independent compact subinterval of $(0, +\infty)$. Let us define the following constants β, γ

$$\beta = 1 - 2\eta \in (0,1), \quad \gamma = \frac{1-2\eta}{1+2\eta} \in (0,1). \tag{25}$$

We will show that

$$\beta \le I_1 \le 1/\beta \tag{26}$$

$$\gamma\beta \le I_2 \le 1/\beta \tag{27}$$

Note that for any $\ell \in [m]$, if we define $u_t = \|\Pi_{\mathcal{B}}(\boldsymbol{w}_t)\|^2$ then

$$\min\{1-\eta u_t, 1+\eta(2-u_t-\|\boldsymbol{w}_t\|^2)\} \le \frac{c_{t+1}(\ell)}{c_t(\ell)} \le \max\{1-\eta u_t, 1+\eta(2-u_t-\|\boldsymbol{w}_t\|^2)\}.$$

By Proposition 15 since $0 \le u_t \le \|\boldsymbol{w}_t\|^2 \le 2$,

$$\min\{1-\eta u_t, 1+\eta(2-u_t-\|\boldsymbol{w}_t\|^2)\} \ge 1-2\eta$$
$$\max\{1-\eta u_t, 1+\eta(2-u_t-\|\boldsymbol{w}_t\|^2)\} \le \max\{1, 1+2\eta\} = 1+2\eta.$$

Therefore for β defined as in Eq. (25)

$$\frac{c_{t+1}(\ell)}{c_t(\ell)} \in [\beta, 1/\beta]$$

for any $\ell \in [m]$. From this it easily follows $I_1 \in [\beta, 1/\beta]$ as claimed in Eq. (26) since

$$\frac{c_{t+1}(i^\star_{t+1})}{c_t(i^\star_t)} \le \frac{c_{t+1}(i^\star_{t+1})}{c_t(i^\star_{t+1})} \le 1/\beta, \quad \frac{c_{t+1}(i^\star_{t+1})}{c_t(i^\star_t)} \ge \frac{c_{t+1}(i^\star_t)}{c_t(i^\star_t)} \ge \beta.$$

Now let us consider the term I_2. If $i^\star_{t+1} = i^\star_t$ then it is easy to see that

$$I_2 = \frac{\sum\limits_{\ell\in\mathcal{J}_t} c_t(\ell)}{\sum\limits_{\ell\in\mathcal{J}_{t+1}} c_{t+1}(\ell)} = \frac{\sum\limits_{\ell\in\mathcal{J}_t} c_t(\ell)}{\sum\limits_{\ell\in\mathcal{J}_t} c_{t+1}(\ell)} \in [\beta, 1/\beta]. \tag{28}$$

Now consider the case when $i^\star_{t+1} \ne i^\star_t$. We claim that is can only happen if $i^\star_{t+1} \in \mathcal{B}_t$ and $i_t \notin \mathcal{B}_t$. To see why, let $A = c_{t+1}(i^\star_{t+1})/c_t(i^\star_t)$ and $B = c_{t+1}(i^\star_t)/c_t(i^\star_t)$. If this were not true then

$$\frac{c_{t+1}(i^\star_{t+1})}{c_{t+1}(i^\star_t)} = \frac{c_t(i^\star_{t+1})}{c_t(i^\star_t)}\frac{A}{B} \le \frac{c_t(i^\star_{t+1})}{c_t(i^\star_t)} \le 1$$

because $A, B \in \{1+\eta(2-u_t-\|\boldsymbol{w}_t\|^2), 1-\eta u_t\}$ and $A/B > 1$ only if $A = 1+\eta(2-u_t-\|\boldsymbol{w}_t\|^2)$ and $B = 1-\eta u_t$, which is exactly when $i^\star_{t+1} \in \mathcal{B}_t$ and $i_t \notin \mathcal{B}_t$. In this case

$$1 \leq \frac{c_{t+1}(i^\star_{t+1})}{c_{t+1}(i^\star_t)} = \frac{c_t(i^\star_{t+1})}{c_t(i^\star_t)} \frac{1+\eta(2-u_t-\|\boldsymbol{w}_t\|^2)}{1-\eta u_t} \leq \frac{c_t(i^\star_{t+1})}{c_t(i^\star_t)} \frac{1+2\eta}{1-2\eta}$$

from which it follows that

$$c_t(i^\star_{t+1}) \geq c_t(i^\star_t)\frac{1-2\eta}{1+2\eta} = \gamma c_t(i^\star_t).$$

Now observe that I_2 lies in the interval

$$I_2 = \frac{\sum\limits_{\ell\in\mathcal{J}_t} c_t(\ell)}{\sum\limits_{\ell\in\mathcal{J}_{t+1}} c_{t+1}(\ell)} \in [\beta I_2', (1/\beta)\cdot I_2'], \tag{29}$$

where we define the term

$$I_2' := \frac{\sum\limits_{\ell\in\mathcal{J}_t} c_t(\ell)}{\sum\limits_{\ell\in\mathcal{J}_{t+1}} c_t(\ell)} = \frac{\sum\limits_{\ell\in[m]\setminus\{i^\star_t, i^\star_{t+1}\}} c_t(\ell) + c_t(i^\star_{t+1})}{\sum\limits_{\ell\in[m]\setminus\{i^\star_t, i^\star_{t+1}\}} c_t(\ell) + c_t(i^\star_t)}.$$

It is clear that $I_2' \leq 1$ because $c_t(i^\star_t) \geq c_t(i^\star_{t+1})$ and since $c_t(i^\star_{t+1}) \geq \gamma c_t(i^\star_t)$ with $\gamma \in (0,1)$,

$$\begin{aligned} I_2' &= \frac{\sum\limits_{\ell\in[m]\setminus\{i^\star_t, i^\star_{t+1}\}} c_t(\ell) + c_t(i^\star_{t+1})}{\sum\limits_{\ell\in[m]\setminus\{i^\star_t, i^\star_{t+1}\}} c_t(\ell) + c_t(i^\star_t)} \\ &\geq \frac{\sum\limits_{\ell\in[m]\setminus\{i^\star_t, i^\star_{t+1}\}} c_t(\ell) + \gamma c_t(i^\star_t)}{\sum\limits_{\ell\in[m]\setminus\{i^\star_t, i^\star_{t+1}\}} c_t(\ell) + c_t(i^\star_t)} \geq \gamma. \end{aligned}$$

Thus we see that $I_2' \in [\gamma, 1]$. From Eq. (29) it follows that in this case $I_2 \in [\gamma\beta, 1/\beta]$. Combining with Eq. (28) yields the desired bound on I_2 in Eq. (27). Since the bounds on I_1 and I_2 imply that $I \in [\gamma\beta^2, 1/\beta^2]$ since this implies

$$|R_{t+1} - R_t| \leq \max(|\log(\gamma\beta^2)|, |\log(1/\beta^2)|).$$

C.2 Proof of Theorem 2

Now we are ready to give the convergence results for minibatch SGD in Theorem 2.

Proof From Proposition 9 we know that $R_t \to \infty$ almost surely, that is the ratio of $|c_t(i^\star_t)|$ to $S_t := \sum_{\ell\in\mathcal{J}_t} |c_t(\ell)|$ goes to infinity, where $\mathcal{J}_t := [m] \setminus \{i^\star_t\}$. Since $|c_t(i^\star_t)| \leq 1$, we know that $S_t \to 0$. We now show that $|c_t(i^\star_t)| \to 1$. With Ψ_t as defined as in Eq. (19) and using $|c_t(\ell)| < 1$ for all $\ell \in [n]$, we have

$$\begin{aligned} 1 \geq c_t(i^\star_t)^2 &= \|\boldsymbol{w}_t\|^2 - \sum_{\ell\in\mathcal{J}_t} c_t(\ell)^2 - \Psi_t \\ &\geq \|\boldsymbol{w}_t\|^2 - |c_t(i^\star_t)|\exp(-R_t) - \Psi_t \\ &\geq \|\boldsymbol{w}_t\|^2 - \exp(-R_t) - \Psi_t. \end{aligned}$$

Now since $\liminf_{t\to\infty} \|\boldsymbol{w}_t\| \geq 1$ by Proposition 22 and $\Psi_t \to 0$ by Proposition 21, since R_t is transient we can see by taking $t \to \infty$ that indeed $|c_t(i_t^\star)| \to 1$.

Now it remains to show that eventually $i_t^\star$ becomes constant. Intuitively, this is true because the fact that $c_t(i_t^\star) \to 1$ while $\max_{\ell\neq i} c_t(\ell)^2 \to 0$ means that the only way for $i_t^\star$ to be non-constant would be for gradient descent to rapidly move all of the mass from one coordinate to another, but this is not possible since the gradient norm goes to zero as we show in Lemma 25. More formally, for any $\varepsilon > 0$, take $T := T(\varepsilon)$ large enough such that $\max_{\ell\neq i_t^\star} |c_t(\ell)| \leq \varepsilon$ and $|c_t(i_t^\star) - 1| \leq \varepsilon$ for all $t \geq T$. By Lemma 25,

$$\sup_{t\geq T} \max_{\ell\in[n]} |c_{t+1}(\ell) - c_t(\ell)| = O(\varepsilon).$$

Let $i^\star := i_T^\star$. We will prove that for ε small enough, $i_t^\star = i^\star$ for all $t \geq T$. For the sake of contradiction, let $t > T$ be the first time such that $i_t^\star \neq i^\star$. From the above we must have

$$\begin{aligned} \max_{\ell\neq i^\star} c_t(\ell) &\leq \max_{\ell\neq i^\star} c_{t-1}(\ell) + O(\varepsilon) \leq O(\varepsilon) \\ c_t(i^\star) &\geq c_{t-1}(i^\star) - O(\varepsilon) \geq 1 - O(\varepsilon) \end{aligned}$$

Therefore for ε small enough $i^\star = \arg\max_{\ell\in[m]} |c_t(\ell)|$ which is a contradiction. Thus we have $i_t^\star = i^\star$ for $t \geq T$ and so $|c_t(i^\star)| \to 1$. Furthermore, by Corollary 16 we have that $\operatorname{sign}(c_t(i^\star)) = \operatorname{sign}(c_0(i^\star))$, hence $\boldsymbol{w}_t \to \operatorname{sign}(c_0(i^\star)) \cdot \boldsymbol{a}_i$ which is what we wished to show. ■

Lemma 25 *Consider a trajectory $(\boldsymbol{w}_t)_{t\in\mathbb{Z}^+}$ of minibatch GD. Assume that for some t, there exists $i \in [m]$ and $\varepsilon > 0$, such that*

$$|c_t(i) - 1| \leq \varepsilon \text{ and } |c_t(j)| \leq \varepsilon \text{ for all } j \in [n] \setminus \{i\}.$$

Then as $\varepsilon \to 0^+$,

$$\max_{\ell\in[n]} |c_{t+1}(\ell) - c_t(\ell)| \leq O(\varepsilon).$$

Proof Define $u_t = \|\Pi_{\mathcal{B}_t}(\boldsymbol{w}_t)\|^2$. Consider $\ell \notin \mathcal{B}_t$, then

$$|c_{t+1}(\ell) - c_t(\ell)| = \eta|c_t(\ell)|u_t \leq \eta m\varepsilon = O(\varepsilon).$$

Now consider $\ell \in \mathcal{B}_t$. Then for ε small enough,

$$\begin{aligned} 2 - u_t - \|\boldsymbol{w}_t\|^2 &\geq 2 - 2\|\boldsymbol{w}_t\|^2 \\ &\geq 2 - 2(1+\varepsilon)^2 - 2(n-1)\varepsilon^2 \\ &= -4\varepsilon - 2n\varepsilon^2, \\ 2 - u_t - \|\boldsymbol{w}_t\|^2 &\leq 2 - (1-\varepsilon)^2 - (1-\varepsilon)^2 \\ &= 4\varepsilon - 2\varepsilon^2. \end{aligned}$$

Hence $|2 - u_t - \|\boldsymbol{w}_t\|^2| = O(\varepsilon)$ and

$$|c_{t+1}(\ell) - c_t(\ell)| = \eta|c_t(\ell)||(2 - u_t - \|\boldsymbol{w}_t\|^2)| = O(\varepsilon).$$

■

Appendix D. Full-batch Gradient Descent Convergence

In this section we will give the proof of Theorem 1 which gives the convergence of full-batch gradient descent. As a reminder, we will make Assumption 1 throughout, that is we assume that $\|\boldsymbol{w}_0\| < 1$ and $\eta \leq 1/5$, where $\eta := \alpha/m$. Let us recall some definitions. Define the vector of correlations

$$\boldsymbol{c}_t = (\langle \boldsymbol{w}_t, \boldsymbol{a}_1\rangle, \dots, \langle \boldsymbol{w}_t, \boldsymbol{a}_n\rangle) \in \mathbb{R}^n$$

Furthermore we will define

$$\Phi_t := \|\Pi_m(\boldsymbol{w}_t)\|^2 = \sum_{i\in[m]} c_t(i)^2, \qquad \Psi_t := \|\boldsymbol{w}_t\|^2 - \|\Pi_m(\boldsymbol{w}_t)\|^2 = \sum_{j\in[n]\setminus[m]} c_t(j)^2.$$

From Lemma 14 we can write the full-batch gradient update as follows

$$\begin{aligned} c_{t+1}(i) &= c_t(i) + \eta c_t(i)(2 - 2\Phi_t - \Psi_t) && i \in [m],\\ c_{t+1}(j) &= c_t(j) - \eta c_t(j)\Phi_t && j \in [n]\setminus[m]. \end{aligned}$$

To reduce notational clutter in the following we will sometimes suppress the time index t and write for example $c_i := c_t(i)$, $c_i' := c_{t+1}(i)$, and $\Delta c_i = c_i' - c_i$. For example, we can write the full-batch update at time t as follows

$$\Delta c_i = \eta c_i(2 - 2\Phi - \Psi) \qquad i \in [m], \tag{30}$$

$$\Delta c_j = -\eta c_j \Phi \qquad j \in [n]\setminus[m]. \tag{31}$$

We will first show that the dynamics obey an important invariant due to the symmetry of the updates which is not true when the batch size $b < m$.

Proposition 26 *For all $i \in [m]$ and for all $t \in \mathbb{Z}^+$,*

$$\frac{c_t(i)}{c_0(i)} = \sqrt{\frac{\Phi_t}{\Phi_0}}.$$

Proof Define the quantity,

$$\Gamma_t = \eta(2 - 2\Phi_t - \Psi_t).$$

From Eq. (30)

$$\Delta c_i = \eta c_i \Gamma, \quad i \in [m].$$

or equivalently

$$c_{t+1}(i) = c_t(i)(1 + \Gamma_t), \quad i \in [m].$$

Unrolling the updates over t yields

$$c_t(i) = c_0(i)\prod_{k=0}^{t-1}(1+\Gamma_k), \qquad i \in [m].$$

By squaring and summing both sides of the above over $i \in [m]$ we see

$$\Phi_t = \Phi_0\left[\prod_{k=0}^{t-1}(1+\Gamma_k)\right]^2$$

It immediately follows that

$$\frac{c_t(i)}{c_0(i)} = \sqrt{\frac{\Phi_t}{\Phi_0}}, \qquad i \in [m].$$

■

Therefore to obtain the convergence behavior of $c_t(i)$ for $i \in [m]$ it suffices to understand the limit of Φ_t. The same proof given in Proposition 21 will give $\Psi_t \to 0$ for full-batch GD, which implies $c_t(j) \to 0$ for all $j \notin [n] \setminus [m]$, but this fact will emerge in the proofs from this section anyways. Thus, to analyse the convergence of $\boldsymbol{w}_t$ it suffices to analyse the limits of Φ_t and Ψ_t. We can easily write the update equations for the dynamics of Φ_t and Ψ_t solely in terms of these two quantities.

Lemma 27 *The updates for Φ_t and Ψ_t are given by*

$$\begin{aligned}
\Delta\Phi &= 2\eta\Phi(2 - 2\Phi - \Psi) + \eta^2\Phi(2 - 2\Phi - \Psi)^2 \\
\Delta\Psi &= -2\eta\Phi\Psi + \eta^2\Phi^2\Psi.
\end{aligned}$$

Proof This follow from straight-forward calculations

$$\begin{aligned}
\Delta\Phi = \Phi' - \Phi &= \sum_{i \in S} (c_i')^2 - c_i^2 \\
&= \sum_{i \in S} (c_i' - c_i)(c_i' + c_i) \\
&= \sum_{i \in S} \eta c_i(2 - 2\Phi - \Psi)(2c_i + \eta c_i(2 - 2\Phi - \Psi)) \\
&= 2\eta \sum_{i \in S} c_i^2(2 - 2\Phi - \Psi) + \eta^2 \sum_{i \in S} c_i^2(2 - 2\Phi - \Psi)^2 \\
&= 2\eta\Phi(2 - 2\Phi - \Psi) + \eta^2\Phi(2 - 2\Phi - \Psi)^2,
\end{aligned}$$

and similarly

$$\begin{aligned}
\Delta\Psi = \Psi' - \Psi &= \sum_{j \in S^c} (c_j')^2 - c_j^2 \\
&= \sum_{j \in S^c} (c_j' - c_j)(c_j' + c_j) \\
&= \sum_{j \in S^c} -\eta c_j \Phi(2c_j - \eta c_j \Phi) \\
&= -2\eta \sum_{j \in S^c} c_j^2 \Phi + \eta^2 \sum_{j \in S^c} c_j^2 \Phi^2 \\
&= -2\eta\Phi\Psi + \eta^2\Phi^2\Psi.
\end{aligned}$$

■

The next proposition establishes the asymptotic convergence of Φ_t and Ψ_t. In particular we will show the following

Proposition 28 *Φ_t monotonically increases to 1 and Ψ_t monotonically decreases to 0.*

To show the above proposition, we will first show that the following quantity

$$\mathcal{N}_t = \Phi_t + \frac{5}{8}\Psi_t$$

remains bounded above by one, from which the proposition will easily follow.

Lemma 29 *If $\mathcal{N}_t < 1$, then $\mathcal{N}_{t+1} < 1$.*

Proof Consider the update at time t. By definition

$$\Delta\mathcal{N} = \Delta\Phi + \frac{5}{8}\Delta\Psi.$$

From Lemma 27 we have that

$$\begin{aligned}\Delta\Phi &= 2\eta\Phi(2-2\Phi-\Psi) + \eta^2\Phi(2-2\Phi-\Psi)^2\\ \Delta\Psi &= -2\eta\Phi\Psi + \eta^2\Phi^2\Psi.\end{aligned}$$

We will show that

$$\begin{aligned}\Delta\Phi &\le 5\eta\Phi(1-\Phi-\Psi/2),\\ \Delta\Psi &\le -\eta\Phi\Psi.\end{aligned}$$

Since $\Phi + (5/8)\Psi < 1$ and $\Phi, \Psi \ge 0$ it follows that

$$0 \le \Phi < 1 \text{ and } 0 < 1-\Phi-\Psi/2 \le 1.$$

Furthermore since $\eta \le 1/5$, we can bound $\Delta\Phi$ as follows

$$\begin{aligned}\Delta\Phi &= 2\eta\Phi(2-2\Phi-\Psi) + \eta^2\Phi(2-2\Phi-\Psi)^2\\ &= 4\eta\Phi(1-\Phi-\Psi/2) + 4\eta(1-\Phi-\Psi/2)\cdot[\eta\Phi(1-\Phi-\Psi/2)]\\ &\le 4\eta\Phi(1-\Phi-\Psi/2) + \frac{4}{5}\eta(1-\Phi-\Psi/2)\\ &\le 5\eta\Phi(1-\Phi-\Psi/2).\end{aligned}$$

Similarly, for $\Delta\Psi$ we have

$$\begin{aligned}\Delta\Psi &= -2\eta\Phi\Psi + \eta^2\Phi^2\Psi\\ &= -2\eta\Phi\Psi + \eta\Phi\Psi[\eta\Phi]\\ &\le -2\eta\Phi\Psi + \frac{1}{5}\eta\Phi\Psi \le -\eta\Phi\Psi.\end{aligned}$$

Now observe that from the previous inequalities

$$\begin{aligned}\Delta\mathcal{N} &= \Delta\Phi + \frac{5}{8}\Delta\Psi\\ &\le 5\eta\Phi(1-\Phi-\Psi/2) - \frac{5}{8}\eta\Phi\Psi\\ &= 5\eta\Phi(1-(\Phi+5\Psi/8))\\ &\le 5\eta(\Phi+5\Psi/8)(1-(\Phi+5\Psi/8))\\ &= 5\eta\mathcal{N}(1-\mathcal{N}).\end{aligned}$$

Since $\eta \leq 1/5$ by Assumption 1, it follows $\mathcal{N}_{t+1} < 1$ by a simple calculation (see Lemma 31). ■

We are now ready to give the proof of Proposition 28.

Proof [Proposition 28] We will first show that $\Phi_t \to 1$. Since $\mathcal{N}_0 < 1$ by Assumption 1, from Lemma 29 it follows by induction that $\mathcal{N}_t = \Phi_t + (5/8)\Psi_t < 1$ for all t. Let us now consider the updates at a particular time. We have that

$$2 - 2\Phi - \Psi \geq 2(1-\Phi) - \frac{8}{5}(1-\Phi) = \frac{2}{5}(1-\Phi) > 0.$$

Recall from Lemma 27 that

$$\begin{aligned} \Delta\Phi &= 2\eta\Phi(2 - 2\Phi - \Psi) + \eta^2\Phi(2 - 2\Phi - \Psi)^2 \\ &\geq 2\eta\Phi(2 - 2\Phi - \Psi). \end{aligned}$$

Thus we see that since $\Delta\Phi \geq 0$, Φ_t is monotonically increasing and

$$\begin{aligned} \Delta\Phi &\geq 2\eta\Phi(2 - 2\Phi - \Psi) \\ &\geq (2\eta) \cdot \frac{2}{5}(1-\Phi) \\ &= \frac{4}{5}\eta\Phi(1-\Phi) \\ &\geq \frac{4}{5}\eta\Phi_0 \cdot (1-\Phi). \end{aligned}$$

Thus, by a simple calculation (see Lemma 30), for all t we have

$$0 \leq 1 - \Phi_t \leq (1-\Phi_0) \cdot \exp(-\kappa t)$$

where $\kappa := (4/5)\eta\Phi_0 > 0$, hence $\Phi_t \to 1$ as desired. From Corollary 16 we know Ψ_t is monotonically decreasing. By the squeeze theorem it is easy to see that $\Psi_t \to 0$, since $\Phi_t + (5/8)\Psi_t < 1$ implies that

$$0 \leq \Psi_t \leq \frac{8}{5}(1-\Phi_t).$$

■

Now we are ready to give the proof of Theorem 1.

Proof [Theorem 1] Recall from Proposition 26 that

$$\frac{c_t(i)}{c_0(i)} = \sqrt{\frac{\Phi_t}{\Phi_0}}$$

for all $i \in [m]$ and $t \in \mathbb{Z}^+$. From Proposition 28 we have $\Phi_t \to 1$, hence

$$c_t(i) \to \frac{c_0(i)}{\sqrt{\Phi_0}}, \qquad i \in [m],$$

as well as $\Psi_t \to 0$, from which it is clear that $c_t(j) \to 0$ for $j \in [n] \setminus [m]$. Therefore we see

$$\boldsymbol{w}_t \to \frac{1}{\sqrt{\Phi_0}} \sum_{i \in [m]} c_0(i)\boldsymbol{a}_i$$

as we wished to show. ■

We now give the proof of Corollary 3. For convenience, we will say an event occurs with high probability (w.h.p) if it occurs with probability at least $1 - O(m^{-1})$.

Proof [Corollary 3] First note that therefore by Theorem 2, it follows that

Consider the GD iterates $\boldsymbol{w}_t$. Define the normalized iterates $\overline{\boldsymbol{w}}_t = \boldsymbol{w}_t / \|\boldsymbol{w}_t\|$. By Theorem 1 we have that on the event $\|\boldsymbol{w}_0\| < 1$,

$$\lim_{t\to\infty} \overline{\boldsymbol{w}}_t = \frac{\Pi_m(\boldsymbol{w}_0)}{\|\Pi_m(\boldsymbol{w}_0)\|}, \qquad \Pi_m(\boldsymbol{w}_0) = \sum_{i\in[m]} \langle \boldsymbol{w}_0, \boldsymbol{a}_i\rangle \boldsymbol{a}_i,$$

and so

$$\begin{aligned} \operatorname{cossim}(\boldsymbol{w}_t^{\mathrm{GD}}, \mathcal{D}) &= \max_{i\in[m]} \ |\langle \boldsymbol{a}_i, \lim_{t\to\infty} \overline{\boldsymbol{w}}_t\rangle| \ = \frac{1}{\|\Pi_m(\boldsymbol{w}_0)\|} \cdot \max_{i\in[m]} |c_0(i)|, \\ \operatorname{cossim}(\boldsymbol{w}_t^{\mathrm{GD}}, \boldsymbol{w}_0) &= |\langle \overline{\boldsymbol{w}}_0, \lim_{t\to\infty} \overline{\boldsymbol{w}}_t\rangle| \ = \frac{\|\Pi_m(\boldsymbol{w}_0)\|}{\|\boldsymbol{w}_0\|}. \end{aligned}$$

Note that since $c_0(i) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\text{init}}^2/n)$ it follows that

$$\|\Pi_m(\boldsymbol{w}_0)\|^2 = \sum_{i\in[m]} c_0(i)^2 \sim \frac{\sigma_{\text{init}}^2}{n} \cdot \chi^2(m) \ \text{ and } \ \|\boldsymbol{w}_0\|^2 = \sum_{i\in[n]} c_0(i)^2 \sim \frac{\sigma_{\text{init}}^2}{n} \cdot \chi^2(n),$$

where $\chi^2(k)$ denotes a chi-squared random variable with k degrees of freedom. The standard tail bound in Lemma 32 implies that w.h.p,

$$\sigma_{\text{init}}^2 \cdot \frac{m}{4n} \le \|\Pi_m(\boldsymbol{w}_0)\|^2 \le \sigma_{\text{init}}^2 \cdot \frac{4m}{n}, \qquad \frac{\sigma_{\text{init}}^2}{4} \le \|\boldsymbol{w}_0\|^2 \le \sigma_{\text{init}} < 1.$$

Furthermore by a standard inequality for the maximum absolute value of independent Gaussians stated in Lemma 34, we have that w.h.p,

$$\max_{i\in[m]} |c_0(i)| \ \le 3\sigma_{\text{init}} \sqrt{\frac{\log(2m)}{n}}. \tag{32}$$

Thus it is easy to see that w.h.p

$$\begin{aligned} \operatorname{cossim}(\boldsymbol{w}_t^{\mathrm{GD}}, \mathcal{D}) &= \frac{1}{\|\Pi_m(\boldsymbol{w}_0)\|} \cdot \max_{i\in[m]} |c_0(i)| = O\left(\sqrt{\frac{\log m}{m}}\right), \\ \operatorname{cossim}(\boldsymbol{w}_t^{\mathrm{GD}}, \boldsymbol{w}_0) &= \frac{\|\Pi_m(\boldsymbol{w}_0)\|}{\|\boldsymbol{w}_0\|} = \Theta\Big(\frac{m}{n}\Big). \end{aligned}$$

Now if $\boldsymbol{w}_t$ are the iterates of SGD, then by Theorem 2, on the event $\|\boldsymbol{w}_0\| < 1$, it holds that almost surely

$$\lim_{t\to\infty} \overline{\boldsymbol{w}}_t \in \{s \cdot \boldsymbol{a}_i : i \in [m], s \in \{-1, +1\}\}.$$

Hence by Eq. (32) it is clear that w.h.p

$$\begin{aligned}\operatorname{cossim}(\boldsymbol{w}_t^{\text{SGD}}, \mathcal{D}) &= 1,\\ \operatorname{cossim}(\boldsymbol{w}_t^{\text{SGD}}, \boldsymbol{w}_0) &= \max_{i\in[m]} |c_0(i)| = O\left(\sqrt{\frac{\log m}{n}}\right).\end{aligned}$$

■

D.1 Technical Lemmas

Lemma 30 *Consider a sequence $\{x_t\}_{t\in\mathbb{N}}$ which satisfies*

$$x_{t+1} - x_t \geq c_t(1 - x_t)$$

for all $t \in \mathbb{N}$, where $c_t \in (0, 1]$ and $x_0 \leq 1$. Then

$$1 - x_t \leq \prod_{i=1}^{t}(1 - c_i)(1 - x_0) \leq \exp\left(-\sum_{i=1}^{t} c_i\right)(1 - x_0)$$

Proof Rearranging

$$x_{t+1} - x_t \geq c_t(1 - x_t)$$

yields

$$(1 - x_{t+1}) \leq (1 - c_t)(1 - x_t)$$

hence unrolling the recursion yields

$$1 - x_t \leq \prod_{i=1}^{t}(1 - c_i)(1 - x_0)$$

and then the inequality $1 - x \leq e^{-x}$ yields

$$\prod_{i=1}^{t}(1 - c_t)(1 - x_0) \leq \exp\left(-\sum_{i=1}^{t} c_i\right)(1 - x_0).$$

■

Lemma 31 *Let $\{x_t\}_{t\in\mathbb{N}}$ be a sequence such that $x_0 < 1$ and*

$$x_{t+1} - x_t \leq \lambda x_t(1 - x_t)$$

for $\lambda \leq 1$. Then $x_t < 1$ for all $t \in \mathbb{N}$.

Proof Assume the statement is true for $t \leq T$. Observe that the function

$$f(x) = (1+\lambda)x - \lambda x^2$$

has derivative

$$f'(x) = 1 + \lambda - 2\lambda x$$

hence f is strictly increasing on the interval $(-\infty, 1]$ and $f(1) = 1$. Therefore since $x_T \in [0,1)$, we have that $x_{T+1} \leq f(x_T) < 1$ completing the claim. ■

D.2 Concentration Inequalities

Lemma 32 (Chi-square Tail Bound) *If $X \sim \chi^2(k)$ then for all $t \in (0,1)$,*

$$\Pr[X \leq k(1-t)] \leq \exp\bigl(-kt^2/8\bigr).$$

Lemma 33 (Chernoff Bound) *Let $X = \sum_{i=1}^n X_i$ where $X_i = 1$ with probability p_i and $X_i = 0$ with probability $1 - p_i$, and all X_i are independent. Let $\mu = \mathbb{E}(X) = \sum_{i=1}^n p_i$. Then*

$$\Pr(X \leq (1-\delta)\mu) \leq \exp\bigl(-\mu\delta^2/2\bigr)$$

for all $\delta \in (0,1)$.

Lemma 34 (Maximum of Gaussians) *Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. Then,*

$$\Pr\left(\max_{i \in [n]} |X_i| - \sqrt{2\sigma^2 \log(2n)} \geq t\right) \leq \exp\left(\frac{-t^2}{2\sigma^2}\right).$$

Appendix E. Cyclic SGD Convergence

E.1 Dynamics of Cyclic SGD

First let us recall the setting of Theorem 4. We assume that $m = n = 2$, that is our dataset $\mathcal{D} = \{\boldsymbol{a}_0, \boldsymbol{a}_1\}$ where $\boldsymbol{a}_i \in \mathbb{R}^2$. At each time step t we process example $\boldsymbol{x}_t$ where $\boldsymbol{x}_t = \boldsymbol{a}_{t\%2}$ and $t\%2$ is 0 when t is even and 1 when t is odd. Let $y_t := \langle \boldsymbol{w}_t, \boldsymbol{a}_0 \rangle$ and $z_t := \langle \boldsymbol{w}_t, \boldsymbol{a}_1 \rangle$. We assume that $y_0 \geq z_0 > 0$. From Lemma 14 it follows that the dynamics are given by

$$\begin{aligned} y_{t+1} &= y_t(1 + \eta(2 - 2y_t^2 - z_t^2)) \\ z_{t+1} &= z_t(1 - \eta y_t^2) \end{aligned}$$

for $t\%2 = 0$ and for $t\%2 = 1$

$$\begin{aligned} y_{t+1} &= y_t(1 - \eta z_t^2) \\ z_{t+1} &= z_t(1 + \eta(2 - 2z_t^2 - y_t^2)). \end{aligned}$$

For convenience we let $F : \mathbb{R}^2 \to \mathbb{R}^2$ denote the function which gives the two-step (epoch) update

$$(y_{t+2}, z_{t+2}) = F(y_t, z_t), \quad t\%2 = 0.$$

Note that by Proposition 15 and 16, if $y_0^2 + z_0^2 < 1$ and $\eta \leq 1/4$ then for all t,

$$(y_t, z_t) \in \{(y, z) : 0 < y, z < 1, \quad y^2 + z^2 \leq 1 + \eta/4\}.$$

We will make use of the following definitions

- Define the potential function $V(y, z) = z/y$, which gives the relative alignment (compare with Eq. (11)). We will show that the potential is always decreasing by at least a constant each epoch in order to prove that $V(y_t, z_t) \to 0$.

- Define $\mathcal{V}_- = \{(y, z) \in (0, 1)^2 : V(F(y, z)) - V(y, z) < 0\}$ as the set of points where the potential strictly decreases after an epoch. We will show V decreases each epoch by proving that the iterates are always in this set (in particular the subset $\mathcal{A}$ defined below) at the start of the epoch.

- Define $\mathcal{A} = \{(y, z) \in (0, 1)^2 : y \geq z > 0, y^2 + z^2 \leq 1 + \eta/4\}$. We will show that $\mathcal{A} \subseteq \mathcal{V}_-$ and that $\mathcal{A}$ is an invariant set under F, i.e. $(y, z) \in \mathcal{A}$ implies $F(y, z) \in \mathcal{A}$.

E.2 Proof of Theorem 4

We will consider the subsequence of even iterates (y_{2t}, z_{2t}) for $t = 0, 1, \ldots$ Let us recall the sets $\mathcal{V}_-$ and $\mathcal{A}$ and the epoch update function F defined in Appendix E.1. In Proposition 36 we show that $\mathcal{A} \subseteq \mathcal{V}_-$. By the definition of $\mathcal{V}_-$ it is easy to see that $\mathcal{A}$ is invariant under F, that is if $(y, z) \in \mathcal{A}$ then $F(y, z) \in \mathcal{A}$. Since by assumption $(y_0, z_0) \in \mathcal{A}$, this will imply that $(y_{2t}, z_{2t}) \in \mathcal{A}$ for all t and that $V(y_{2t}, z_{2t})$ is strictly decreasing. Thus by the monotone convergence theorem there exists some $V_\star \in [0, +\infty)$ such that $V(y_{2t}, z_{2t}) \to V_\star$.

We claim that $V_\star = 0$. For the sake of contradiction assume that $V_\star > 0$. Let $N_t = y_t^2 + z_t^2$. Since by assumption $N_0 < 1$, from Lemmas 17, 18 and Proposition 15, we have that for all t,

$$0 < N_0 \leq N_{2t} \leq 1 + \eta/4.$$

Since $V_\star \leq z_{2t}/y_{2t} \leq z_0/y_0 \leq 1$, the sequence (y_{2t}, z_{2t}) lies in the annulus $\mathcal{K}_1$ where

$$\mathcal{K}_1 = \{(r\cos\theta, r\sin\theta) : \tan(\theta) \in [V_\star, 1], r \in [N_0, 1 + \eta/4]\}.$$

Note that $\mathcal{K}_1 \subseteq \mathcal{V}_-$ is a compact set. Thus we have a contradiction by Proposition 35 and therefore $V_\star = 0$. Now we show that $\lim(y_{2t}, z_{2t}) = (1, 0)$. By Corollary 16

$$y_{2t}^2 = N_{2t} - (z_{2t}/y_{2t})^2 \cdot y_{2t}^2 \geq N_{2t} - (z_{2t}/y_{2t})^2$$

which implies that

$$\begin{aligned} \liminf y_{2t}^2 &\geq \liminf N_{2t} - \lim(z_{2t}/y_{2t})^2 \\ &= \liminf N_{2t} - V_\star \\ &= \liminf N_{2t} \geq 1, \end{aligned}$$

where the last inequality follows from Proposition 22 because $i_{2t} = i_{2t}^\star = 0$ for all t. Since $y_{2t}^2 \leq 1$ we have $\limsup y_{2t}^2 \leq 1$. Therefore $\lim y_{2t} = 1$ and $\lim z_{2t} = \lim y_{2t} \cdot (z_{2t}/y_{2t}) = 0$. We have shown that the even subsequence converges to the desired limit point. Now invoking Lemma 25 is is easy to see that $(y_t, z_t) \to (1, 0)$ as desired.

E.3 Auxiliary Results

Proposition 35 *Let $\{x_t\}_{t=0}^{\infty}$ be a sequence in $\mathbb{R}^n$ such that there exists continuous $F : \mathbb{R}^n \to \mathbb{R}^n$ and $x_{t+1} = F(x_t)$ for all $t = 0, 1 \dots$ Assume there exists a function $V : \mathbb{R}^n \to \mathbb{R}$ that is continuous on a compact subset $\mathcal{K} \subseteq \mathbb{R}^n$ such that for all $x \in \mathcal{K}$, $V(F(x)) - V(x) < 0$. Then there exists $t_0 \in \mathbb{N}$ such that $x_{t_0} \notin \mathcal{K}$.*

Proof For the sake of contradiction assume that $x_t \in \mathcal{K}$ for all t. Define the quantity

$$\varepsilon := \sup\{V(F(x)) - V(x) : x \in \mathcal{K}\}.$$

By the continuity of V and F and the compactness of $\mathcal{K}$, it follows that $\varepsilon < 0$. Therefore for any T,

$$\begin{aligned}\inf_{x\in\mathcal{K}} V(x) &\leq V(x_T)\\ &= V(x_0) + \sum_{t=0}^{T-1} V(x_{t+1}) - V(x_t)\\ &= V(x_0) + \sum_{t=0}^{T-1} V(F(x_t)) - V(x_t)\\ &\leq V(x_0) + \varepsilon T.\end{aligned}$$

However, the inequality

$$\inf_{x\in\mathcal{K}} V(x) \leq V(x_0) + \varepsilon T$$

cannot hold since the left-hand side is finite and the right-hand side approaches negative infinity as $T \to \infty$. ■

Proposition 36 *The set $\mathcal{A} = \{(y, z) : y \geq z > 0, y^2 + z^2 \leq 1 + \eta/4\} \subseteq \mathcal{V}_-$.*

Proof Let $y_0 = r\cos\theta$ and $z_0 = r\sin\theta$ with $\theta \in [0, \pi/2]$. Consider fixing r and varying θ. Observe that

$$V(F(r\cos\theta, r\sin\theta)) - V(r\cos\theta, r\sin\theta) = \tan\theta \cdot \left(\frac{(1-\eta y_0^2)}{(1+\eta(2-2y_0^2-z_0^2))}\frac{(1+\eta(2-2z_1^2-y_1^2)}{(1-\eta z_1^2)} - 1\right).$$

Therefore $(y_0, z_0) \in \mathcal{V}_-$ iffthe following inequality holds

$$\frac{(1-\eta y_0^2)}{(1+\eta(2-2y_0^2-z_0^2))} \leq \frac{(1-\eta z_1^2)}{(1+\eta(2-2z_1^2-y_1^2)}$$

or equivalently

$$(1-\eta y_0^2)(1+\eta(2-2z_1^2-y_1^2) \leq (1-\eta z_1^2)(1+\eta(2-2y_0^2-z_0^2)).$$

Let us observe that we can write the following terms solely as a function of r and y_0.

$$\begin{aligned}
z_0^2 &= r - y_0^2 \\
z_1^2 &= z_0^2(1-\eta y_0^2)^2 = (r-y_0^2)(1-\eta y_0^2)^2 \\
y_1^2 &= y_0^2(1+\eta(2-2y_0^2-z_0^2)^2 = y_0^2(1+\eta(2-r-y_0))^2.
\end{aligned}$$

Letting $y = y_0$ for convenience and substituting into the above inequality, it is equivalent to

$$f(y;r) - g(y;r) \leq 0$$

where

$$\begin{aligned}
f(y;r) &= (1-\eta y^2)(1+\eta[2-2(r-y^2)(1-\eta y^2)^2 - y^2(1+\eta(2-r-y^2))^2]) \\
g(y;r) &= (1-\eta(r-y^2)(1-\eta y)^2)(1+\eta(2-r-y^2)).
\end{aligned}$$

By Lemma 39

$$\frac{\mathrm{d}}{\mathrm{d}y} f(y;r) - g(y;r) \leq 0.$$

Recalling $y = r\cos\theta$, by the chain rule

$$\frac{\mathrm{d}}{\mathrm{d}\theta}[f(y(\theta);r) - g(y(\theta);r)] = \frac{\mathrm{d}}{\mathrm{d}y}[f(y;r)-g(y;r)]\frac{\mathrm{d}y}{\mathrm{d}\theta} = \frac{\mathrm{d}}{\mathrm{d}y}[f(y;r)-g(y;r)](-r\sin\theta) \geq 0. \tag{33}$$

As $\cos(\pi/4) = \sin(\pi/4) = 1/\sqrt{2}$, Lemma 37 states that if $r \leq \sqrt{1+\eta/4}$ then $(r\cos\pi/4, r\sin\pi/4) \in \mathcal{V}_-$, that is

$$f(r\cos\pi/4;r) - g(r\cos\pi/4;r) < 0.$$

From Eq. (33) for $0 \leq \psi \leq \pi/4$

$$f(r\cos\psi;r) - g(r\cos\psi;r) \leq f(r\cos\pi/4;r) - g(r\cos\pi/4;r) < 0$$

hence $(r\cos\psi, r\sin\psi) \in \mathcal{V}_-$. Since

$$\mathcal{A} = \{(r\cos\psi, r\sin\psi) : r^2 \leq 1+\eta/4, \psi \in [0,\pi/4]\}$$

this proves the claim. ■

Lemma 37 *If* $0 < y^2 \leq \frac{1}{2}(1+\eta/4)$ *and* $\eta \leq 1/4$*, then* $(y,y) \in \mathcal{V}_-$*.*

Proof Observe that

$$(y,y) \in \mathcal{V}_- \iff \frac{z_2}{y_2} - 1 > 0 \iff y_2 - z_2 > 0.$$

We will explicitly show that the last inequality for y such that $y^2 \leq (1+\eta/4)/2$. We have that

$$\begin{aligned}
y_1 &= y(1+\eta(2-3y^2)) \\
z_1 &= y(1-\eta y^2)
\end{aligned}$$

Therefore

$$y_1 = (1+\delta)z_1, \qquad \delta = \frac{2\eta(1-y^2)}{1-\eta y^2}.$$

Thus we have that

$$\begin{aligned} y_2 - z_2 &= y_1(1-\eta z_1^2) - z_1(1+\eta(2-2z_1^2-y_1^2)) \\ &= \delta z_1 - \eta(1+\delta)z_1^3 - 2\eta z_1 + 2\eta z_1^3 + \eta(1+\delta)^2 z_1^3 \\ &= z_1(\delta - 2\eta) + \eta z_1^3(2+\delta+\delta^2) \end{aligned}$$

Substituting and factoring yields

$$z_1(\delta-2\eta)+\eta z_1^3(2+\delta+\delta^2) = 2\eta z_1 y^2\left(\frac{(\eta-1)}{1-\eta y^2} + (1-\eta y^2)^2\left(1+\frac{\eta(1-y^2)}{1-\eta y^2}+\frac{2\eta^2(1-y^2)^2}{(1-\eta y^2)^2}\right)\right)$$

Letting $w = 1-\eta y^2$ we thus $y_2 - z_2 \geq 0$ iff

$$\frac{\eta-1}{w} + w^2\left(1+\frac{\eta(1-y^2)}{w}+\frac{2\eta^2(1-y^2)^2}{w^2}\right) > 0$$

Letting $b = 1-\eta$, it follows that $\eta(1-y^2) = w-b$ and so the above is equivalent to

$$-\frac{b}{w} + w^2\left(1+\frac{w-b}{w}+\frac{2(w-b)^2}{w^2}\right) > 0$$

which after clearing denominators and grouping terms is equivalent to

$$4w^3 - 5bw^2 + 2wb^2 - b > 0.$$

The claim then follows from Lemma 38. ■

E.4 Technical Lemmas

Lemma 38 *Assume $\eta \leq 1/4$ and $2y^2 \leq 1+\eta/4$. Let $w = 1-\eta y^2$ and $b = 1-\eta$. Then*

$$4w^3 - 5bw^2 + 2wb^2 - b \geq 0.$$

Proof Let $f(w,b) = 4w^3 - 5bw^2 + 2wb^2 - b$. Since by assumption $y^2 \leq (1+\eta/4)/2$,

$$w \geq 1 - \eta/2 - \eta^2/8 = \frac{1}{8}(-b^2+6b+3).$$

Let us call

$$w_{\min} = \frac{1}{8}(-b^2+6b+3).$$

Observe that for $w \in [b,1]$

$$\frac{\mathrm{d}}{\mathrm{d}w} f(w,b) = 12w^2 - 10bw + 2b^2 \geq 14b^2 - 10b \geq 0$$

since

$$14b^2 - 10b \geq 0 \iff b \geq 5/7 \iff \eta \leq 2/7$$

which is true since by assumption $\eta \leq 1/4 \leq 2/7$. Further, note that $w_{\min} \geq b$ since

$$8(w_{\min} - b) = -b^2 + 6b + 3 - 8b = -(b^2 + 2b - 3) = -(b+3)(b-1),$$

and $8(w_{\min} - b) > 0$ for $b \in [0, 1]$. We thus have,

$$\inf_{w \in [w_{\min}, 1]} f(w, b) = f(w_{\min}, b).$$

Using Mathematica to simplify

$$f(w_{\min}, b) = -\frac{1}{128}(b-1)^2(b^4 - 6b^3 - 2b^2 + 2b - 27).$$

Since $b \in [0, 1)$

$$b^4 - 6b^3 - 2b^2 + 2b - 27 \leq 1 + 2 - 27 < 0$$

hence $f(w_{\min}, b) > 0$. ■

Lemma 39 *Assume $r \leq 1 + \eta/4$ is a constant. Define the following functions of $y \in [0, 1]$.*

$$\begin{aligned} f(y; r) &= (1 - \eta y^2)(1 + \eta[2 - 2(r - y^2)(1 - \eta y^2)^2 - y^2(1 + \eta(2 - r - y^2))^2]) \\ g(y; r) &= (1 - \eta(r - y^2)(1 - \eta y)^2)(1 + \eta(2 - r - y^2)). \end{aligned}$$

Then the following is true

$$\frac{\mathrm{d}}{\mathrm{d}y} f(y; r) - g(y; r) \leq 0$$

Proof Making the substitution $w = 1 - \eta y^2 \iff \eta y^2 = 1 - w$ we have

$$\begin{aligned} f(w; r) &= (1 - \eta y^2)(1 + \eta[2 - 2(r - y^2)(1 - \eta y^2)^2 - y^2(1 + \eta(2 - r - y^2))^2]) \\ &= (1 - \eta y^2)(1 + 2\eta + 2(\eta y^2 - \eta r)(1 - \eta y^2)^2 - \eta y^2(1 - \eta y^2 + \eta(2 - r))^2]) \\ &= w[1 + 2\eta + 2w^2(1 - w - \eta r) + (w - 1)(w + \eta(2 - r))^2]. \\ g(w; r) &= (1 - \eta(r - y^2)(1 - \eta y)^2)(1 + \eta(2 - r - y^2)) \\ &= (1 + (\eta y^2 - \eta r)(1 - \eta y)^2)(1 - \eta y^2 + \eta(2 - r)) \\ &= (1 + w^2(1 - w - \eta r))(w + \eta(2 - r)). \end{aligned}$$

Using Mathematica we have that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}w} f(w; r) - g(w; r) &= \eta[\eta(2 - r)(r - 2 + 4w) + 6(3 - 2r)w^2 - 6(2 - r)w + 2] \\ &= \eta[p(r, w) + q(r, w)]. \end{aligned}$$

where

$$\begin{aligned} p(r, w) &= \eta(2 - r)(r - 2 + 4w) \\ q(r, w) &= 6(3 - 2r)w^2 + 6(r - 2)w + 2. \end{aligned}$$

We now show that $p(r, w) \geq 0$ and $q(r, w) \geq 0$.

Proof that $p(r, w) \geq 0$
Note that $r \leq 1 + \eta/4 \leq 2$ hence $2 - r \geq 0$ and since $y^2 \leq 1$ it follows that $w \geq 1 - \eta$, hence $4w \geq 4(1 - \eta) \geq 3$ hence $(r - 2 + 4w) \geq r + 1 \geq 0$ since $r \geq 0$. Therefore $p(r, w) = \eta(2 - r)(r - 2 + 4w) \geq 0$.

Proof that $q(r, w) \geq 0$
Note that we can write

$$q(r, w) = 6(3 - 2r)w^2 + 6(r - 2)w + 2 = 6rw(1 - 2w) + s(w)$$

for some function s of w. Since $1 - 2w \leq 1 - 2(1 - \eta) = -1 + 2\eta \leq 0$ it follows that q is decreasing in r therefore $q(r, w) \geq q(1 + \eta/4) \geq q(1 + \eta, w)$. We can lower bound this as follows, using $\eta \leq 1/4$

$$\begin{aligned} q(1 + \eta, w) &= 6(3 - 2(1 + \eta))w^2 - 6(1 - \eta)w + 2 \\ &\geq 3w^2 - \frac{9}{2}w + 2 \\ &\geq 3(3/4)^2 - (9/2)(3/4) + 2 = 5/16 \geq 0. \end{aligned}$$

Therefore we have shown that

$$\frac{\mathrm{d}}{\mathrm{d}w} f(w; r) - g(w; r) \geq 0$$

and since $w = 1 - \eta y^2$ by the chain rule this implies that

$$\frac{\mathrm{d}}{\mathrm{d}y} f(y; r) - g(y; r) \leq 0.$$

■

Appendix F. Loss Landscape

F.1 Proof of Theorem 12

It is clear that for minimizing the loss objective we can restrict our consideration to $\boldsymbol{w} \in \operatorname{span}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m)$. Let us define $c_i := \langle \boldsymbol{w}, \boldsymbol{a}_i \rangle$ for $i \in [m]$. Then

$$\begin{aligned}
\mathcal{L}(\boldsymbol{w}; \mathcal{D}) &= \frac{1}{2m} \sum_{i=1}^{m} \| \boldsymbol{a}_i - \boldsymbol{w}\phi(\langle \boldsymbol{w}, \boldsymbol{a}_i \rangle) \|^2 \\
&= \frac{1}{2m} \sum_{i=1}^{m} \left\| \boldsymbol{a}_i - \phi(c_i) \sum_{j=1}^{m} c_j \boldsymbol{a}_j \right\|^2 \\
&= \frac{1}{2m} \left(\sum_{i=1}^{m} (1 - c_i \phi(c_i))^2 + \sum_{j \neq i} c_j^2 \phi(c_i)^2 \right).
\end{aligned}$$

Define the quantity $B := \sum_{j=1}^{m} c_j^2$. Then recalling that the activation $\phi(t) = \max(t, 0)$

$$\begin{aligned}
\mathcal{L}(\boldsymbol{w}; \mathcal{D}) &= \frac{1}{2m} \left(\sum_{i=1}^{m} (1 - c_i \phi(c_i))^2 + \phi(c_i)^2 (B - c_i^2) \right) \\
&= \frac{1}{2m} \left(m - 2 \sum_{i=1}^{m} c_i \phi(c_i) + \sum_{i=1}^{m} c_i^2 \phi(c_i)^2 + \sum_{i=1}^{m} \phi(c_i)^2 (B - c_i^2) \right) \\
&= \frac{1}{2m} \left(m - 2 \sum_{i=1}^{m} c_i \phi(c_i) + B \sum_{i=1}^{m} \phi(c_i)^2 \right) \\
&= \frac{1}{2m} \left(m - 2 \sum_{i=1}^{m} c_i \phi(c_i) + \sum_{i=1}^{m} c_i^2 \sum_{i=1}^{m} \phi(c_i)^2 \right).
\end{aligned}$$

Therefore to find a minimizer it suffices to minimize the quantity

$$-2 \sum_{i=1}^{m} c_i \phi(c_i) + \sum_{i=1}^{m} c_i^2 \sum_{i=1}^{m} \phi(c_i)^2. \tag{34}$$

If we define

$$P = \sum_{i: c_i > 0} c_i^2, \qquad N = \sum_{i: c_i < 0} c_i^2.$$

then Eq. (34) can be rewritten as

$$-2P + P(P + N) = P^2 - 2P + PN$$

where $P, N \geq 0$. It is easy to see that the minimum of this quantity is achieved precisely when $P = 1$, $N = 0$, which is what we wished to prove.

F.2 One-sided Derivatives

Due to the presence of the ReLU activation in the auto-encoder (see Eq. (1)), the loss objective $\mathcal{L}(\boldsymbol{w})$ in Eq. (2) is not smooth everywhere since ReLU is not differentiable at 0. The objective is in fact first-order differentiable everywhere due to the squaring, but not second-order differentiable at $\boldsymbol{w}$ such that $\langle \boldsymbol{w}, \boldsymbol{a}_i \rangle = 0$ for some $i \in [m]$. However, just as the ReLU function is one-sided differentiable everywhere, the loss objective has one-sided derivatives of all order everywhere. We will now introduce our notation for one-sided derivatives and related definitions.

Given a function $f : \mathbb{R}^n \to \mathbb{R}$ and a direction $\boldsymbol{v} \in \mathbb{R}^n$, define the one-sided derivative of f in the direction v at a point $\boldsymbol{x} \in \mathbb{R}^n$ as the following scalar quantity

$$D_{\boldsymbol{v}}\ f(\boldsymbol{x}) := \lim_{t \to 0^+} \frac{f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x})}{t}.$$

We can define the second directional derivative analogously as

$$D^2_{\boldsymbol{v}}\ f(\boldsymbol{x}) = \lim_{t \to 0^+} \frac{D_{\boldsymbol{v}}\ f(\boldsymbol{x} + t\boldsymbol{v}) - D_{\boldsymbol{v}}\ f(\boldsymbol{x})}{t}.$$

Note that if f is first-order differentiable at $\boldsymbol{x}$ then

$$D_{\boldsymbol{v}}\ f(\boldsymbol{x}) = \langle \boldsymbol{\nabla}_{\boldsymbol{x}}\ f(\boldsymbol{x}), \boldsymbol{v} \rangle,$$

and if f is twice differentiable then

$$D^2_{\boldsymbol{v}}\ f(\boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{\nabla}^2_{\boldsymbol{x}}\ f(\boldsymbol{x})\boldsymbol{v}.$$

We will use these notions to define measures of sharpness which generalize the standard Hessian based measures.

F.3 Generalized Sharpness

In the literature, the two prevalent notions of sharpness of a function $\boldsymbol{x}$ at a point $\boldsymbol{x}$ are the maximum eigenvalue and the trace of the Hessian of f at $\boldsymbol{x}$. That is, if f is twice-differentiable and $\boldsymbol{H}(\boldsymbol{x}) := \boldsymbol{\nabla}^2_{\boldsymbol{x}}\ f(\boldsymbol{x})$ has eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n$, then the two measures can be written as

1. (Maximum Curvature) $\|\boldsymbol{H}(\boldsymbol{x})\|_2 = \lambda_1$,
2. (Average Curvature) $\mathrm{Tr}(\boldsymbol{H}(\boldsymbol{x})) = \sum_{i=1}^n \lambda_i$.

The first quantity measures the curvature in the maximal direction. The second quantity can be seen to be a measure of average curvature over random directions since by a well-known identity

$$\mathrm{Tr}(\boldsymbol{H}(\boldsymbol{x})) = \mathbb{E}_{\boldsymbol{v} \sim \mathcal{N}(0,\mathbf{I})}\ \boldsymbol{v}^\top \boldsymbol{H}(\boldsymbol{x})\boldsymbol{v}.$$

For both measures, large values indicate increased sharpness. We now introduce sharpness measures which generalize the previous ones, but are well-defined for functions with only one-sided derivatives

1. (Maximum Curvature) $\left\|D^2 f(\boldsymbol{x})\right\|_2 := \sup_{\|\boldsymbol{v}\|=1} D^2_{\boldsymbol{v}}\ f(\boldsymbol{x})$,
2. (Average Curvature) $\mathrm{Tr}\big(D^2 f(\boldsymbol{x})\big) := \mathbb{E}_{\boldsymbol{v}\sim\mathcal{N}(0,\mathbf{I})}\ D^2_{\boldsymbol{v}}\ f(\boldsymbol{x})$.

From the previous section, we know that these measures are in fact generalizations, since if f is differentiable then

$$\left\|D^2 f(\boldsymbol{x})\right\|_2 = \|\boldsymbol{H}(\boldsymbol{x})\|_2, \quad \mathrm{Tr}\big(D^2 f(\boldsymbol{x})\big) = \mathrm{Tr}(\boldsymbol{H}(\boldsymbol{x})).$$

F.4 Sharpness at Global Minima

Using the sharpness measures defined in the previous section, we now explicitly compute the sharpness of the loss objective $\mathcal{L}(\boldsymbol{w})$ at the convergence points of GD and (C)SGD.

First note that we can rewrite the loss objective as

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{m}\sum_{i\in[m]} f_i(\boldsymbol{w}) + \mathrm{const}, \quad f_i(\boldsymbol{w}) = \phi^2(\langle \boldsymbol{w}, \boldsymbol{a}_i\rangle)\cdot(\|\boldsymbol{w}\|^2/2 - 1). \tag{35}$$

Indeed by expanding the square in Eq. (2)

$$\begin{aligned}
\mathcal{L}(\boldsymbol{w}) &= \frac{1}{2m}\sum_{i\in[m]} \|\boldsymbol{a}_i - \boldsymbol{w}\phi(\langle \boldsymbol{w}, \boldsymbol{a}_i\rangle)\|^2 \\
&= \frac{1}{m}\sum_{i\in[m]} \frac{1}{2}\|\boldsymbol{a}_i\|^2 - \langle \boldsymbol{w}, \boldsymbol{a}_i\rangle\phi(\langle \boldsymbol{w}, \boldsymbol{a}_i\rangle) + \frac{1}{2}\|\boldsymbol{w}\|^2\phi^2(\langle \boldsymbol{w}, \boldsymbol{a}_i\rangle) \\
&= \frac{1}{2} + \frac{1}{m}\sum_{i\in[m]} -\phi(\langle \boldsymbol{w}, \boldsymbol{a}_i\rangle)^2 + \frac{1}{2}\|\boldsymbol{w}\|^2\phi^2(\langle \boldsymbol{w}, \boldsymbol{a}_i\rangle) \\
&= \frac{1}{2} + \frac{1}{m}\sum_{i\in[m]} \phi^2(\langle \boldsymbol{w}, \boldsymbol{a}_i\rangle)\cdot(\|\boldsymbol{w}\|^2/2 - 1).
\end{aligned}$$

Therefore we can write the second directional derivative of $\mathcal{L}$ as

$$D^2_{\boldsymbol{v}}\ \mathcal{L}(\boldsymbol{w}) = \frac{1}{m}\sum_{i\in[m]} D^2_{\boldsymbol{v}}\ f_i(\boldsymbol{w}). \tag{36}$$

Observe that if we define for $i \in [m]$,

$$g_i(\boldsymbol{w}) = \langle \boldsymbol{w}, \boldsymbol{a}_i\rangle^2\cdot(\|\boldsymbol{w}\|^2/2 - 1),$$

then the second derivative of f_i at $\boldsymbol{w}$ is

$$D^2_{\boldsymbol{v}}\ f_i(\boldsymbol{w}) = \begin{cases} D^2_{\boldsymbol{v}}\ g_i(\boldsymbol{w}) & \text{if } \langle \boldsymbol{w}, \boldsymbol{a}_i\rangle \geq 0 \text{ and } \langle \boldsymbol{v}, \boldsymbol{a}_i\rangle \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{37}$$

Since $g_i(\boldsymbol{w})$ is twice-differentiable we can compute the gradient and Hessian as

$$\begin{aligned}
\boldsymbol{\nabla}_{\boldsymbol{w}}\ g_i(\boldsymbol{w}) &= \langle \boldsymbol{w}, \boldsymbol{a}_i\rangle(\|\boldsymbol{w}\|^2 - 2)\boldsymbol{a}_i + \langle \boldsymbol{w}, \boldsymbol{a}_i\rangle^2\boldsymbol{w}, \\
\boldsymbol{\nabla}^2_{\boldsymbol{w}}\ g_i(\boldsymbol{w}) &= (\|\boldsymbol{w}\|^2 - 2)\boldsymbol{a}_i\boldsymbol{a}_i^\top + 2\langle \boldsymbol{w}, \boldsymbol{a}_i\rangle(\boldsymbol{a}_i\boldsymbol{w}^\top + \boldsymbol{w}\boldsymbol{a}_i^\top) + \langle \boldsymbol{w}, \boldsymbol{a}_i\rangle^2\cdot\mathbf{I}_n.
\end{aligned} \tag{38}$$

For a set $\mathcal{S} \subseteq [m]$, consider points

$$\boldsymbol{w} = \sum_{i \in \mathcal{S}} c_i \boldsymbol{a}_i \text{ such that } c_i > 0 \text{ for all } i \in \mathcal{S} \text{ and } \sum_{i \in \mathcal{S}} c_i^2 = 1.$$

We will be interested in such points because our convergence theorems show that GD converges to the point $\boldsymbol{w}_{\mathrm{GD}}$ (see Eq. (16)) for which $\mathcal{S} = \mathcal{S}^+$ and SGD converges to the point $\boldsymbol{w}_{\mathrm{SGD}}$ (see Eq. (17)) for which $|\mathcal{S}| = 1$.

F.5 Proof of Theorem 13

Computing Maximum Curvature Note that if $\langle \boldsymbol{w}, \boldsymbol{a}_i \rangle = 0$ and $\langle \boldsymbol{v}, \boldsymbol{a}_i \rangle > 0$ then

$$\boldsymbol{\nabla}_{\boldsymbol{w}}^2 \, g_i(\boldsymbol{w}) = (\|\boldsymbol{w}\|^2 - 2)\boldsymbol{a}_i \boldsymbol{a}_i^\top = -\boldsymbol{a}_i \boldsymbol{a}_i^\top,$$

which is negative semidefinite. Hence for $i \notin \mathcal{S}$, $D_{\boldsymbol{v}}^2 f_i(\boldsymbol{w}) \geq D_{\boldsymbol{v}}^2 g_i(\boldsymbol{w})$. Therefore if we are trying to maximize $D_{\boldsymbol{v}}^2 \mathcal{L}(\boldsymbol{w})$ with respect to $\boldsymbol{v}$, we can restrict our consideration to $\boldsymbol{v}$ such that $\langle \boldsymbol{v}, \boldsymbol{a}_i \rangle = 0$ if $i \notin \mathcal{S}$. For such $\boldsymbol{v}$, by Eq. (36) and (37)

$$\begin{aligned}
D_{\boldsymbol{v}}^2 \, \mathcal{L}(\boldsymbol{w}) &= \frac{1}{m} \sum_{i \in \mathcal{S}} D_{\boldsymbol{v}}^2 \, g_i(\boldsymbol{w}) \\
&= \frac{1}{m} \boldsymbol{v}^\top \left(\sum_{i \in \mathcal{S}} \boldsymbol{\nabla}_{\boldsymbol{w}}^2 \, g_i(\boldsymbol{w}) \right) \boldsymbol{v} \\
&= \frac{1}{m} \boldsymbol{v}^\top \left(\sum_{i \in \mathcal{S}} 2\langle \boldsymbol{w}, \boldsymbol{a}_i \rangle (\boldsymbol{a}_i \boldsymbol{w}^\top + \boldsymbol{w} \boldsymbol{a}_i^\top) \right) \boldsymbol{v} + \frac{1}{m} \boldsymbol{v}^\top \left(\sum_{i \in \mathcal{S}} -\boldsymbol{a}_i \boldsymbol{a}_i^\top + \langle \boldsymbol{w}, \boldsymbol{a}_i \rangle^2 \mathbf{I}_n \right) \boldsymbol{v} \\
&= \frac{1}{m} \boldsymbol{v}^\top \left(\sum_{i \in \mathcal{S}} 2\langle \boldsymbol{w}, \boldsymbol{a}_i \rangle (\boldsymbol{a}_i \boldsymbol{w}^\top + \boldsymbol{w} \boldsymbol{a}_i^\top) \right) \boldsymbol{v} \\
&= \frac{4}{m} \sum_{i \in \mathcal{S}} \langle \boldsymbol{w}, \boldsymbol{a}_i \rangle \langle \boldsymbol{v}, \boldsymbol{a}_i \rangle \langle \boldsymbol{w}, \boldsymbol{v} \rangle \\
&= \frac{4}{m} \langle \boldsymbol{w}, \boldsymbol{v} \rangle^2.
\end{aligned}$$

From this it is easy to see that $\boldsymbol{v} = \boldsymbol{w}$ is a maximizer and $\left\| D^2 \mathcal{L}(\boldsymbol{w}) \right\|_2 = 4/m$.

Computing Average Curvature Now let us compute $\mathrm{Tr}\big(D^2 \mathcal{L}(\boldsymbol{w})\big)$. Observe that from Eq. (37)

$$\begin{aligned}
\mathrm{Tr}\big(D^2 f_i(\boldsymbol{w})\big) &= \mathrm{Tr}\big(D^2 g_i(\boldsymbol{w})\big) = \mathrm{Tr}\big(\boldsymbol{\nabla}_{\boldsymbol{w}}^2 \, g_i(\boldsymbol{w})\big) && \text{if } \langle \boldsymbol{w}, \boldsymbol{a}_i \rangle > 0 \\
\mathrm{Tr}\big(D^2 f_i(\boldsymbol{w})\big) &= \frac{1}{2} \mathrm{Tr}\big(D^2 g_i(\boldsymbol{w})\big) = \frac{1}{2} \mathrm{Tr}\big(\boldsymbol{\nabla}_{\boldsymbol{w}}^2 \, g_i(\boldsymbol{w})\big) && \text{if } \langle \boldsymbol{w}, \boldsymbol{a}_i \rangle = 0
\end{aligned}$$

where the second line is due to the fact that $\Pr_{\boldsymbol{v} \sim \mathcal{N}(0,\mathbf{I})}(\langle \boldsymbol{v}, \boldsymbol{a}_i \rangle \leq 0) = 1/2$. Note that

$$\mathrm{Tr}\big(\boldsymbol{\nabla}_{\boldsymbol{w}}^2 \, g_i(\boldsymbol{w})\big) = (\|\boldsymbol{w}\|^2 - 2) + (n+4)\langle \boldsymbol{w}, \boldsymbol{a}_i \rangle^2.$$

and that by Eq. (36) and the linearity of $\mathrm{Tr}\big(D^2 \bullet\big)$

$$\mathrm{Tr}\big(D^2\mathcal{L}(\boldsymbol{w})\big) = \frac{1}{m}\sum_{i\in[m]} \mathrm{Tr}\big(D^2 f_i(\boldsymbol{w})\big).$$

Therefore by Eq. (37)

$$\begin{aligned}
\mathrm{Tr}\big(D^2\mathcal{L}(\boldsymbol{w})\big) &= \frac{1}{m}\left[\sum_{\ell\in\mathcal{S}} \mathrm{Tr}\big(\boldsymbol{\nabla}^2_{\boldsymbol{w}}\; g_\ell(\boldsymbol{w})\big) + \sum_{i\in[m]\setminus\mathcal{S}} \mathrm{Tr}\big(\boldsymbol{\nabla}^2_{\boldsymbol{w}}\; g_i(\boldsymbol{w})\big)/2\right] \\
&= \frac{1}{m}\left[\sum_{\ell\in\mathcal{S}^+} -1 + (n+4)c_\ell^2 + \sum_{i\in[m]\setminus\mathcal{S}} (-1/2)\right] \\
&= \frac{1}{m}[-|\mathcal{S}| + (n+4) - (m - |\mathcal{S}|)/2] = \frac{2n+8-m-|\mathcal{S}|}{2m}.
\end{aligned}$$

Thus in particular we see that

$$\begin{aligned}
\mathrm{Tr}\big(D^2\mathcal{L}(\boldsymbol{w}_{\mathrm{GD}})\big) &= \frac{2n+8-m-|\mathcal{S}^+|}{2m}, \\
\mathrm{Tr}\big(D^2\mathcal{L}(\boldsymbol{w}_{\mathrm{SGD}})\big) &= \frac{2n+7-m}{2m}.
\end{aligned}$$