OPT2022: 14th Annual Workshop on Optimization for Machine Learning

Gradient dynamics of single-neuron autoencoders on orthogonal data

Nikhil Ghosh NIKHIL_GHOSH@BERKELEY.EDU
Spencer Frei FREI@BERKELEY.EDU
Wooseok Ha HAYWSE@BERKELEY.EDU
Bin Yu BINYU@BERKELEY.EDU
University of California, Berkeley

Abstract

In this work we investigate the dynamics of (stochastic) gradient descent when training a single-neuron ReLU autoencoder on orthogonal inputs. We show that for this non-convex problem there exists a manifold of global minima all with the same maximum Hessian eigenvalue and that gradient descent reaches a particular global minimum when initialized randomly. Interestingly, which minimum is reached depends heavily on the batch-size. For full batch gradient descent, the directions of the neuron that are initially positively correlated with the data are merely rescaled uniformly, hence in high-dimensions the learned neuron is a near uniform mixture of these directions. On the other hand, with batch-size one the neuron exactly aligns with a single such direction, showing that when using a small batch-size a qualitatively different type of "feature selection" occurs.

1. Introduction and Related Work

Recent years have witnessed the impressive successes of neural networks across a wide variety of domains. However their ability to generalize to unseen data is still not fully understood [16, 28]. One potential explanation is that gradient-based optimization algorithms have an "implicit bias" towards models that can generalize well [2, 3, 8, 9, 12, 17, 18, 20, 22]. In particular, 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 [11, 13, 25]. These observations have inspired a surge of research aimed at more deeply understanding the particular benefits of small-batch stochastic gradient descent (SGD) over full-batch gradient descent (GD) [5, 10, 15, 26]. Most of this prior work has focused on the supervised learning setting where the data is labeled. However, given the currently massive interest in unsupervised learning [4, 6, 19], it is crucial to better understand the implicit bias of optimization algorithms in the unsupervised setting.

In this work, we consider a simple unsupervised setting where we are given a dataset of orthogonal input vectors and train a single-neuron autoencoder to reconstruct the inputs using gradient descent started from a random initialization. Since the network has only a single neuron, it is generally impossible to perfectly reconstruct all of the inputs, but it is still of interest to understand *what* gradient descent will learn, the quality of learned solutions, and the role of different hyperparameters.

In this setting, we show that there exists a manifold of solutions which achieve the global minimum value, and that gradient descent with a random initialization is able to find a minimum. However, for different choices of the batch size, gradient descent finds qualitatively different minima. In the full

batch setting (c.f. Section 2.3), we show that gradient descent essentially only modifies the norm of the random initialization: the direction of the learned weight vector is almost identical to its randomly-initialized direction. In contrast, for batch-size one we observe empirically (c.f. Section 3) that SGD "rotates" the neuron significantly during training, eventually aligning it with a single datapoint, and prove that this occurs in a simplified setting (c.f. Section 2.4). Additionally, we show that the maximum Hessian eigenvalue at these minima are identical, suggesting that this measure of "flatness" is insufficient to characterize the implicit bias in this setting.

We note that previous works have also considered the dynamics of gradient descent for learning single-neuron architectures [7, 14, 23, 27]. However, to the best of our knowledge 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.

2. Main Results

2.1. Setting

Our model of interest is a simple weight-tied auto-encoder $f(\boldsymbol{x}; \boldsymbol{w}) : \mathbb{R}^n \to \mathbb{R}^n$

$$f(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{w}\phi(\langle \boldsymbol{w}, \boldsymbol{x} \rangle), \quad \phi(t) = \max(t, 0) \tag{1}$$

parameterized by one-neuron $\boldsymbol{w} \in \mathbb{R}^n$, with no bias, and ReLU activation. Assume we are given a dataset $\mathcal{D} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m\}$ where the $\boldsymbol{a}_i \in \mathbb{R}^n$ are orthonormal and necessarily $m \leq n$. Let $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$ be the completion 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

$$\mathcal{L}(\boldsymbol{w}; \mathcal{D}) = \frac{1}{m} \sum_{i=1}^{m} \ell(\boldsymbol{w}; \boldsymbol{a}_i), \quad \ell(\boldsymbol{w}; \boldsymbol{x}) = \frac{1}{2} \left\| \boldsymbol{x} - f(\boldsymbol{x}; \boldsymbol{w}) \right\|^2. \tag{2}$$

Remark 1 *One can view this setting as a very simple instance of the sparse coding model popular in the dictionary learning literature (e.g. [1, 21]) where the ground-truth dictionary is orthogonal and the latent codes are one-hat encodings with no observation noise.*

We will consider SGD training with batch-size b and constant step-size α, namely

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \alpha \frac{1}{b} \sum_{i \in \mathcal{B}(t)} \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}; \boldsymbol{a}_i), \quad \mathcal{B}(t) \subseteq [m], \ |\mathcal{B}(t)| = b, \ t = 0, 1, \ldots \tag{3}$$

where a simple calculation gives the gradient of the pointwise loss

$$\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}; \boldsymbol{x}) = \phi'(\langle \boldsymbol{w}, \boldsymbol{x} \rangle) \cdot [\boldsymbol{x}\boldsymbol{w}^\top + \langle \boldsymbol{w}, \boldsymbol{x} \rangle \mathbf{I}_n] \cdot (f(\boldsymbol{x}; \boldsymbol{w}) - \boldsymbol{x}), \quad \phi'(t) := \mathbf{1}(t \geq 0). \tag{4}$$

We will be most interested in understanding the convergence behavior from a random initialization $w_i(0) \sim_{iid} \mathcal{N}(0, \sigma_{\text{init}}^2/n)$, $i \in [n]$ for some constant $\sigma_{\text{init}} > 0$. There are many possible instantiations of Eq. (3) based on the choice of mini-batch order $\mathcal{B}(t)$ including the following

(GD) Full-batch GD ($b = m$) where $\mathcal{B}(t) = [m]$ for all t.

(SGD) Stochastic GD ($b = 1$) where $\mathcal{B}(t) = \{i_t\}$ and $i_t \sim_{iid} \mathsf{Unif}([m])$ for all t.

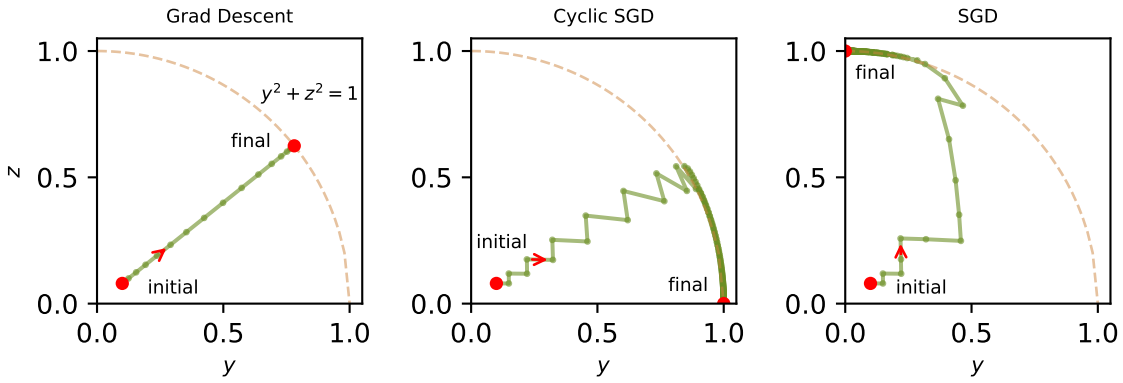

Figure 1: Visualization of optimization trajectories for GD, CSGD, and SGD with $m = n = 2$. All three methods are initialized at $\boldsymbol{w}(0) = (0.1, 0.08)^\top$ and run with step size $\alpha = 0.25$.

(CSGD) Cyclic Stochastic GD ($b = 1$) where $\mathcal{B}(t) = \{t \bmod m\}$.

Our theoretical results analyse in particular GD in Section 2.3 and CSGD in Section 2.4.

2.2. Visualization of convergence behaviors on toy example

To illustrate how the batch size and mini-batch order influence the solutions found by gradient descent, we run full-batch GD, stochastic GD, and cyclic stochastic GD on the simple toy example where $\mathcal{D} = \{\boldsymbol{a}_1, \boldsymbol{a}_2\}$ with $\boldsymbol{a}_1 = (1, 0)^\top$ and $\boldsymbol{a}_2 = (0, 1)^\top$, that is, the dataset is given by standard basis vectors in $\mathbb{R}^2$. For all three methods, we initialize at $\boldsymbol{w}(0) = (0.1, 0.08)^\top$. As we will see in Section 2.3 and Section 2.4, full-batch GD must converge to $\boldsymbol{w}_\star = (0.781, 0.625)$ whereas cyclic SGD converges to $\boldsymbol{a}_1$.

Figure 1 visualizes the optimization trajectory of the coordinates for each of the three methods in $\mathbb{R}^2$. We see that both GD and CSGD converge to points as predicted by our theory. Figure 1 also shows that SGD converges to $\boldsymbol{a}_2$, showing that randomness in the mini-batch order can lead to a different convergence behavior compared with cyclic SGD. Unlike CSGD, SGD is seen with $\boldsymbol{a}_2$ more often during an early stage of iterations, and hence it converges to $\boldsymbol{a}_2$ eventually.

2.3. Full Batch Gradient Descent

We now characterize the dynamics of full-batch gradient descent training. That is we analyse the dynamics of $\boldsymbol{w}(t)$ when $b = m$ in Eq. (3). First let us define the following set

$$S(t) = \{i \in [m] : \langle \boldsymbol{w}_t, \boldsymbol{a}_i \rangle > 0\}, \tag{5}$$

that is $S(t)$ is the indices of datapoints with which $\boldsymbol{w}$ is positively correlated at time t. For convenience let $S := S(0)$. Due to our assumption of random initialization we can assume that $\langle \boldsymbol{w}(0), \boldsymbol{a}_i \rangle \neq 0$ for all $i \in [n]$ since this occurs almost surely. Thus we can assume that S is non-empty, otherwise from Eq. (4) it is easy to see that $\boldsymbol{w}(t) = \boldsymbol{w}(0)$ for all t. Let Π_S be the

orthogonal projection onto $\mathrm{span}(\boldsymbol{a}_i : i \in S)$, that is

$$\Pi_S(\boldsymbol{x}) = \sum_{i \in S} \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle \, \boldsymbol{a}_i.$$

We then have the following limiting characterization the proof of which is given in Appendix A.

Theorem 2 *Assume that the initialization $\boldsymbol{w}(0)$ satisfies the following*

1. $\|\boldsymbol{w}(0)\| < 1$,
2. $|S| > 0$ *and* $\langle \boldsymbol{a}_i, \boldsymbol{w}(0) \rangle \neq 0 \;\; \forall i \in [n]$

and the step-size $\alpha \leq m/5$. Then full-batch gradient descent iterates $\boldsymbol{w}(t)$ converges to $\boldsymbol{w}_\star$ where

$$\boldsymbol{w}_\star = \frac{\Pi_S(\boldsymbol{w}(0))}{\|\Pi_S(\boldsymbol{w}(0))\|}.$$

Corollary 3 *Assume that $\boldsymbol{w}_i(0) \sim_{iid} \mathcal{N}(0, \sigma_{\text{init}}^2/n)$ for $i \in [n]$ where $\sigma_{\text{init}} < 1$ is a constant and $m = \Omega(n)$. Then with probability at least $1 - O(n^{-1})$,*

$$\max_{i \in [m]} \lim_{t \to \infty} \langle \boldsymbol{a}_i, \overline{\boldsymbol{w}}(t) \rangle = \widetilde{O}(n^{-1/2}), \quad \overline{\boldsymbol{w}}(t) = \boldsymbol{w}(t) / \|\boldsymbol{w}(t)\|.$$

2.4. Cyclic Stochastic Gradient Descent

In this section we consider the case of batch-size one. We analyse a simplified set-up where $m = n = 2$ and the mini-batch order remains fixed throughout. For convenience, we will relabel the data indices so that $\mathcal{D} = \{\boldsymbol{a}_0, \boldsymbol{a}_1\}$ and assume that $\mathcal{B}(t) = \{\boldsymbol{a}_{t\%2}\}$ where $t\%2$ is 0 if t is even and 1 if t is odd. In Conjecture 6 we conjecture that similar results hold true for more general settings.

As in Section 2.3 we can assume that at initialization $\langle \boldsymbol{a}_i, \boldsymbol{w}_0 \rangle \neq 0$ for $i \in [n]$. Recall the definition of the set $S := S(0)$ from Eq. (5). From the updates in Eq. (4) it is clear that $\boldsymbol{w}_t = \boldsymbol{w}_0$ for all t if S is empty. If $|S| = 1$, then it is clear that $\boldsymbol{w}_t \to a_i$ where $i \in S$ since the dynamics are equivalent to full-batch gradient descent with batch-size one on the dataset $\mathcal{D} = \{\boldsymbol{a}_i\}$. Therefore, we concentrate on the case when $|S| = 2$. We have the following characterization proven in Appendix B.

Theorem 4 *Assume that $m = n = 2$ and the initialization $\boldsymbol{w}(0)$ satisfies the following*

1. $\langle \boldsymbol{w}(0), \boldsymbol{a}_0 \rangle > 0$ *and* $\langle \boldsymbol{w}(0), \boldsymbol{a}_0 \rangle \geq \langle \boldsymbol{w}(0), \boldsymbol{a}_1 \rangle$
2. $\|\boldsymbol{w}(0)\| < 1$, $\langle \boldsymbol{w}, \boldsymbol{a}_i \rangle \neq 0$ *for all* $i = 0, 1$

and the step-size $\alpha \leq 1/4$. Then the CSGD iterates $\boldsymbol{w}(t)$ converge to $\boldsymbol{a}_0$ as $t \to \infty$.

Corollary 5 *Assume $m = n = 2$ and $w_i(0) \sim_{iid} \mathcal{N}(0, \sigma_{\text{init}}^2/n)$. Then with probability at least some universal constant $\delta > 0$, running full-batch gradient descent and cyclic stochastic gradient descent initialized from $\boldsymbol{w}(0)$ converge to difference solutions.*

Our result in Theorem 4 is limited in the fact that it only covers the case where $m = n = 2$, $\langle \boldsymbol{w}(0), \boldsymbol{a}_0 \rangle \geq \langle \boldsymbol{w}(0), \boldsymbol{a}_1 \rangle$, and the mini-batches follow a fixed, cyclic order. However, we believe that this result can be a useful stepping stone for showing a much more general behavior of SGD which we conjecture below. We have observed this conjectured behavior consistently in simulations (e.g. Section 3) and are currently working on providing a theoretical analysis.

Conjecture 6 *For any $m, n \in \mathbb{N}$ such that $1 \leq m \leq n$ running SGD (or cyclic SGD) on the autoencoder objective Eq. (1) from an initialization $\boldsymbol{w}(0)$ such that $\langle \boldsymbol{w}(0), \boldsymbol{a}_i \rangle > 0$ for some $i \in [m]$ and with step-size $\alpha = O(1)$ will almost surely converge to $\boldsymbol{a}_i$ for some $i \in S$.*

2.5. Loss Landscape

In this section we will study properties of different stationary points of GD and (cyclic) SGD for the one-neuron autoencoder Eq. (1). Our first result characterizes the manifold of global minima.

Theorem 7 (Global Minima) *The minimum value of the loss objective $\mathcal{L}(\boldsymbol{w})$ from Eq. (2) is equal to $\mathcal{L}^\star$ and is attained on the set $\mathcal{M}$ where*

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

Theorem 2 shows that full batch gradient descent converges to the following solution

$$\boldsymbol{w}_\star^{\mathrm{GD}} = \sum_{i \in S} \frac{\langle \boldsymbol{w}_0, \boldsymbol{a}_i \rangle}{\sqrt{\Phi}} \boldsymbol{a}_i, \quad \Phi = \sum_{i \in S} \langle \boldsymbol{w}_0, \boldsymbol{a}_i \rangle^2 \tag{6}$$

where S is defined in Eq. (5). By the above theorem $\boldsymbol{w}_\star^{\mathrm{GD}}$ is a global minimum. In Conjecture 6 we conjecture that in general (C)SGD converges to

$$\boldsymbol{w}_\star^{\mathrm{SGD}} = \boldsymbol{a}_i, \quad \text{for some } i \in S. \tag{7}$$

Interestingly, this point is also a global minimum. Thus, both algorithms optimally minimize the loss objective, but from a "feature learning" perspective achieve qualitatively different solutions, since SGD learns a "pure" datapoint whereas GD learns a "mixture".

As both algorithms converge to global minima, the solutions reached are identical in terms of loss value and have gradient zero. Thus, it is a natural question to understand the second-order behavior of these critical points. The Hessians at the critical points $\boldsymbol{w}_\star^{\mathrm{GD}}$ and $\boldsymbol{w}_\star^{\mathrm{SGD}}$ are given below.

Proposition 8 *The Hessians of the loss objective at $\boldsymbol{w}_\star^{\mathrm{GD}}$ Eq. (6) and $\boldsymbol{w}_\star^{\mathrm{SGD}}$ Eq. (7) are*

$$\boldsymbol{H}_{\mathrm{GD}} := \nabla_{\boldsymbol{w}}^2 \mathcal{L}(\boldsymbol{w}_\star^{\mathrm{GD}}) = -\frac{1}{m} \sum_{\ell \in S} \boldsymbol{a}_\ell \boldsymbol{a}_\ell^\top + \frac{4}{m} \frac{1}{\Phi} \sum_{\ell \in S} \sum_{i \in S} \langle \boldsymbol{w}_0, \boldsymbol{a}_\ell \rangle \langle \boldsymbol{w}_0, \boldsymbol{a}_i \rangle \boldsymbol{a}_i \boldsymbol{a}_\ell^\top + \frac{1}{m} \cdot \mathbf{I}, \tag{8}$$

$$\boldsymbol{H}_{\mathrm{SGD}} := \nabla_{\boldsymbol{w}}^2 \mathcal{L}(\boldsymbol{w}_\star^{\mathrm{SGD}}) = \frac{3}{m} \boldsymbol{a}_i \boldsymbol{a}_i^\top + \frac{1}{m} \cdot \mathbf{I}. \tag{9}$$

A major thread of deep learning research seeks to understand the connection between the "flatness" of local minima, the properties of different optimization algorithms, and generalization performance. Such measures of flatness are usually related to the eigenspectrum of the Hessian. We characterize the Hessian eigenspectra of $\boldsymbol{H}_{\mathrm{GD}}$ and $\boldsymbol{H}_{\mathrm{SGD}}$ below.

Lemma 9 *The Hessian matrix $\boldsymbol{H}_{\mathrm{GD}}$ from Eq. (8) has eigenvalue $4/m$ with multiplicity 1 corresponding to eigenvector $\boldsymbol{w}_\star^{\mathrm{GD}}$, eigenvalue $1/m$ with multiplicity $n - |S|$ and eigenvalue 0 with multiplicity $|S| - 1$.*

Lemma 10 *The Hessian matrix $\boldsymbol{H}_{\mathrm{SGD}}$ from Eq. (9) has eigenvalue $4/m$ with multiplicity 1 corresponding to eigenvector $\boldsymbol{w}_\star^{\mathrm{SGD}}$ and eigenvalue $1/m$ with multiplicity $n-1$.*

From the above we have the following observations. Both $\boldsymbol{H}_{\mathrm{GD}}$ and $\boldsymbol{H}_{\mathrm{SGD}}$ have the same maximum eigenvalue corresponding to the respective solutions $\boldsymbol{w}_\star^{\mathrm{GD}}$ and $\boldsymbol{w}_\star^{\mathrm{SGD}}$. The matrix $\boldsymbol{H}_{\mathrm{GD}}$ is however generally rank deficient for random initializations since $|S| \geq m/4$ with high probability. In contrast, the matrix $\boldsymbol{H}_{\mathrm{SGD}}$ is full-rank. Lastly we can compute the respective traces

$$\mathrm{Tr}(\boldsymbol{H}_{\mathrm{GD}}) = \frac{4+n-|S|}{m}, \quad \mathrm{Tr}(\boldsymbol{H}_{\mathrm{SGD}}) = \frac{3+n}{m}.$$

Note that if $m = \Omega(n)$, then $|S| = \Omega(n)$ with high probability and so $\mathrm{Tr}(\boldsymbol{H}_{\mathrm{GD}}) \ll \mathrm{Tr}(\boldsymbol{H}_{\mathrm{SGD}})$.

3. Numerical Experiments

Here we use simulated data to investigate the convergence behavior of GD, SGD, and CSGD. We fix $n = 100$ and $m \in \{20, 80\}$, corresponding to a small and large m. The dataset $\mathcal{D} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m\}$ is given by columns of an orthonormal matrix drawn at random, and the initialization $\boldsymbol{w}(0)$ is drawn from $\mathcal{N}(0, (\sigma_{\mathrm{init}}^2/n) \cdot \mathbf{I}_n)$ with $\sigma_{\mathrm{init}} = 0.1$. For each of the methods, we run the method for $T = 10^4$ iterations with step-size $\alpha = 0.25$, and repeat for 100 trials. Table 1 shows the maximum correlations between the limit points of each method and datapoints, i.e., $\max_{i \in [m]} \langle \boldsymbol{w}(T), \boldsymbol{a}_i \rangle$, averaged over 100 trials. We observe that cyclic SGD converges to one of the datapoints, as predicted by our theory in the simplified set-up, and similarly for stochastic GD as we conjecture. Proving this is the subject of current ongoing work. Whereas full-batch GD fails to converge to any of the datapoints, and as expected from Corollary 3, the correlation further degrades as m increases.

Settings	full-batch GD	cyclic SGD	ordinary SGD
$m = 20$	0.612 (0.099)	1.0 ($< 10^{-8}$)	1.0 ($< 10^{-8}$)
$m = 80$	0.394 (0.059)	1.0 ($< 10^{-8}$)	1.0 ($< 10^{-8}$)

Table 1: Maximum correlations between limit points of different methods and datapoints in simulated data with $n = 100$, averaged over 100 simulated datasets.

4. Conclusion

In this work, we studied the dynamics of gradient descent for single-neuron autoencoders, showing that gradient descent with a small enough step-size finds a global minimum for this non-convex problem. Different from previous works about learning single-neuron architectures (e.g. [27]), we show that in our setting the choice of batch-size strongly influences the solution found by gradient descent. Although both full batch GD and cyclic SGD reach global minima of the loss objective, the latter becomes highly correlated with a datapoint leading to an arguably more "meaningful" solution. In addition to the obtained loss, the maximal eigenvalue of the Hessians of both solutions are also identical, suggesting that this notion of sharpness is limited for this setting. Looking ahead, it is an exciting direction for future work to extend to more general settings (see Conjecture 6), especially ones involving non-orthogonal data and multi-neuron autoencoders.

References

[1] 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.

[2] 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.

[3] 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.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[5] 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.

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

[7] 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.

[8] 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, 2018.

[9] 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, 2018.

[10] 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.

[11] 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.

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

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

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

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

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

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

[18] Mary Phuong and Christoph H Lampert. The inductive bias of relu networks on orthogonally separable data. In *International Conference on Learning Representations*, 2020.

[19] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[20] 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.

[21] Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Conference on Learning Theory*, pages 37–1. JMLR Workshop and Conference Proceedings, 2012.

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

[23] 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.

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

[25] 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.

[26] 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.

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

[28] 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.

Appendix A. Proofs for Section 2.3

A.1. Proof of Theorem 2

In this section we will give the proof of Theorem 2. Throughout we will make the assumptions given in the theorem statement. Let us make some definitions. Define the vector of correlations with the datapoints

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

and as before define

$$S(t) = \{i \in [m] : c_i(t) > 0\}, \quad S(t)^c = [n] \setminus S(t)$$

where for convenience we let $S := S(0)$ and $S^c := S^c(0)$. Let Π_S be the orthogonal projection onto $\mathrm{span}(\boldsymbol{a}_i : i \in S)$ and Π_{S^c} be the projection onto the complement, that is

$$\Pi_S(\boldsymbol{x}) = \sum_{i \in S} \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle \, \boldsymbol{a}_i, \quad \Pi_{S^c}(\boldsymbol{x}) = \sum_{j \in S^c} \langle \boldsymbol{a}_j, \boldsymbol{x} \rangle \, \boldsymbol{a}_j = x - \Pi_S(\boldsymbol{x}).$$

Furthermore we will define

$$\Phi(t) := \|\Pi_S(\boldsymbol{w}(t))\|^2 = \sum_{i \in S} c_i(t)^2, \quad \Psi(t) := \|\Pi_{S^c}(\boldsymbol{w}(t)\|^2 = \sum_{j \in S^c} c_j(t)^2.$$

Lastly we define the rescaled step-size $\eta := \alpha/m$. By assumption $\eta \leq 1/9$. If at time t, we have $c_i(t) \neq 0$ for all $i \in [n]$, then from Eqs. (3, 4) we can write the full-batch gradient update as follows

$$\begin{aligned} c_i(t+1) &= c_i(t) + \eta c_i(t)(2 - 2\Phi(t) - \Psi(t)), && i \in S(t) \\ c_j(t+1) &= c_j(t) - \eta c_j(t)\Phi(t). && j \in S(t)^c \end{aligned}$$

To reduce notational clutter in the following we will sometimes suppress the time index t and write for example $c_i := c_i(t)$, $c_i' := c_i(t+1)$, and $\Delta c_i = c_i' - c_i$.

Let $P(t)$ be the following statement:

$$\text{for all } i \in [n], \;\; c_i(t) \neq 0 \text{ and } S(t) = S(0)$$

and let $Q(t)$ be the statement that $P(k)$ is true for all $0 \leq k \leq t$. Observe that by assumption $P(0)$ is true. If $P(t)$ is true, then we can write the full-batch update at time t as follows

$$\Delta c_i = \eta c_i(2 - 2\Phi - \Psi), \qquad i \in S \tag{10}$$

$$\Delta c_j = -\eta c_j \Phi. \qquad j \in S^c \tag{11}$$

We will eventually show in Corollary 14 that $P(t)$ is true for all t. We will first show an invariant for a weighted norm-like quantity under the assumption $Q(t)$ holds.

Lemma 11 *If $Q(t)$ is true, then*

$$\Phi(t) + \frac{5}{8}\Psi(t) < 1. \tag{12}$$

Proof We will prove this by induction on t. By assumption the statement is true for $t = 0$. Assuming the statement is true for time $t-1 \geq 0$, we will then show that it is true for time t.

Assume $Q(t)$ is true and consider the update at time $t-1$.

$$\Delta\Phi \leq 5\eta\Phi(1-\Phi-\Psi/2), \tag{13}$$
$$\Delta\Psi \leq -\eta\Phi\Psi. \tag{14}$$

Since $P(t-1)$ is true, by Lemma 12

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

By the induction hypothesis, $\Phi+(5/8)\Psi<1$. Since $\Phi,\Psi\geq 0$, this implies in particular that $\Phi<1$ and $\Phi+\Psi/2<1$. Combined with the fact that $\eta\leq 1/5$ we obtain Eq. (13)

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

Similarly, for Eq. (14)

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

Now observe that from the previous inequalities

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

Since $\eta\leq 1/5$, by Lemma 21 and the induction hypothesis it follows that Eq. (12) is true. ■

Lemma 12 *Assume $P(t)$ is true. We have the following update equations for Φ and Ψ.*

$$\Delta\Phi = 2\eta\Phi(2-2\Phi-\Psi)+\eta^2\Phi(2-2\Phi-\Psi)^2 \tag{15}$$
$$\Delta\Psi = -2\eta\Phi\Psi+\eta^2\Phi^2\Psi. \tag{16}$$

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

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

■

Now we show some important monotonicity properties of the correlations under the assumption of $Q(t)$. As corollaries we will see that $P(t)$ is true for all t and that Φ and Ψ are monotone quantities.

Lemma 13 *Assume $Q(t)$ is true. Then,*

1. $c_i(t+1) > c_i(t)$ *for* $i \in S$,
2. $c_j(t+1)\cdot c_j(t) > 0$ *and* $|c_j(t+1)| < |c_j(t)|$ *for* $j \in S^c$.

Proof We prove this by induction on t. Assume the statement is true for $t - 1 \geq 0$ and consider the update at time t. By the induction hypothesis $c_i(t) > \ldots > c_i(0) > 0$ for $i \in S$. Since $Q(t)$ is true, by Lemma 11 we have that $\Phi + (9/16)\Psi < 1$ which implies in particular that $2\Phi + \Psi < 2$. Therefore since $P(t)$ is true, from Eq. (10) we have $\Delta c_i = \eta c_i(2 - 2\Phi - \Psi) > 0$ for $i \in S$ which implies $c_i(t+1) > c_i(t)$ which shows the first claim. Furthermore, from Eq. (11) we have $c_j' = (1 - \eta\Phi)c_j$. Since $\eta, \Phi \in (0, 1)$, it follows that $0 < 1 - \eta\Phi < 1$ from which we easily get the second claim. ■

Corollary 14 *If $P(0)$ is true, then $P(t)$ is true for all t.*

Proof This is immediate from the previous lemma which gives that if $i \in S$, then $c_i(t) > 0$ for all t and if $j \in S^c \cap [m]$ then $c_j(t) < 0$ for all t. ■

Corollary 15 *Φ is monotone increasing and Ψ is monotone decreasing.*

Proof This is also immediate from the previous lemma which gives that $c_i(t)^2$ is monotone increasing for $i \in S$ and $c_j(t)^2$ is monotone decreasing for $j \in S^c$. ■

Therefore from now on we can assume $P(t)$ is true for all t, hence the dynamics obey the update equations Eqs. (10, 11) and Eqs. (15, 16). Now let us characterize the limiting behaviors of Φ and Ψ.

Lemma 16 *As $t \to \infty$, $\Phi(t) \to 1$ and $\Psi(t) \to 0$.*

Proof For the first claim observe that from Eq. (15)

$$\Delta\Phi = 2\eta\Phi(2 - 2\Phi - \Psi) + \eta^2\Phi(2 - 2\Phi - \Psi)^2 \geq 2\eta\Phi(2 - 2\Phi - \Psi).$$

Furthermore since $\Phi + (9/16)\Psi < 1$ by Lemma 11, we have

$$2 - 2\Phi - \Psi \geq 2(1 - \Phi) - \frac{16}{9}(1 - \Phi) = \frac{2}{9}(1 - \Phi).$$

Therefore since $\Phi(t)$ is increasing by Corollary 15 we have

$$\Delta\Phi \geq \frac{4}{9}\eta\Phi(1 - \Phi) \geq \frac{4}{9}\eta\Phi(0) \cdot (1 - \Phi).$$

Thus, by Lemma 20

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

where $\kappa := (4/9)\eta\Phi(0) > 0$, hence $\Phi(t) \to 1$. Since

$$0 \leq \Psi(t) \leq \frac{16}{9}(1 - \Phi(t))$$

we see $\Psi(t) \to 0$. ■

Now we are ready to complete the proof of Theorem 2. Define the quantity,

$$\Gamma(t) = \eta(2 - 2\Phi(t) - \Psi(t)).$$

Then by unrolling Eq. (10),

$$c_i(t) = c_i(0) \prod_{k=0}^{t-1} (1 + \Gamma(k)), \quad i \in S. \tag{17}$$

Note that we can write Eq. (15) as

$$\frac{\Delta\Phi}{\Phi} = 2\Gamma + \Gamma^2$$

hence unrolling the update over t yields

$$\begin{aligned}\Phi(t) &= \Phi(0)\prod_{k=0}^{t-1}(1+2\Gamma(k)+\Gamma(k)^2)\\ &= \Phi(0)\left[\prod_{k=0}^{t-1}(1+\Gamma(k))\right]^2.\end{aligned}$$

Therefore we have the relation

$$\prod_{k=0}^{t-1}(1+\Gamma(k)) = \sqrt{\frac{\Phi(t)}{\Phi(0)}}.$$

which combined with Eq. (17) implies that

$$\frac{c_i(t)}{\sqrt{\Phi(t)}} = \frac{c_i(0)}{\sqrt{\Phi(0)}}, \quad i \in S.$$

Since $\Phi(t) \to 1$, we get the same result as before

$$c_i(t) \to \frac{c_i(0)}{\sqrt{\Phi(0)}}, \quad i \in S.$$

Since $\Psi \to 0$ it is clear that $c_j(t) \to 0$ for $j \in S^c$. Therefore we see that

$$\boldsymbol{w}(t) \to \frac{\Pi_S(\boldsymbol{w}(0))}{\|\Pi_S(\boldsymbol{w}(0))\|}$$

as desired.

A.2. Proof of Corollary 3

In this section we give the proof of Corollary 3. For convenience, we will say an event occurs w.h.p. if it occurs with probability at least $1 - O(n^{-1})$. By Theorem 2 we have that

$$\lim_{t\to\infty} \overline{\boldsymbol{w}}(t) = \frac{\Pi_S(\boldsymbol{w}(0))}{\|\Pi_S(\boldsymbol{w}(0))\|}, \quad \Pi_S(\boldsymbol{w}(0)) = \sum_{i\in S} \langle \boldsymbol{w}(0), \boldsymbol{a}_i\rangle \, \boldsymbol{a}_i.$$

Therefore our goal is to show that

$$\max_{i\in[m]} \left\langle \boldsymbol{a}_i, \lim_{t\to\infty} \overline{\boldsymbol{w}}(t)\right\rangle = \frac{1}{\sqrt{\Phi}} \max_{i\in S} c_i(0) = \widetilde{O}(n^{-1/2}),$$

w.h.p where $\Phi = \sum_{i\in S} c_i(0)^2$. We will do so by bounding Φ and $\max_{i\in S} c_i(0)$ individually w.h.p. First of all, note that since $|S|$ follows a Binomial distribution Binom(n,1/2), by a Chernoff bound 18, we have $|S| \geq n/4$ w.h.p. Since $c_i(0) = \langle \boldsymbol{w}(0), \boldsymbol{a}_i\rangle \sim_{iid} \mathcal{N}(0, \sigma_{\text{init}}^2/n)$ it follows that conditional on S

$$\Phi = \sum_{i\in S} c_i(0)^2 \sim \sigma_{\text{init}}^2/n \cdot \chi^2(|S|)$$

where $\chi^2(k)$ denotes a chi-squared random variable with k-degrees of freedom. A standard tail bound Lemma 17 implies that w.h.p,

$$\Phi \geq |S| \cdot \frac{\sigma_{\text{init}}^2}{4n}$$

which combined with our bound on $|S|$ yields $\Phi \geq \sigma_{\text{init}}^2/16$ w.h.p. Furthermore, a standard inequality for the maximum of independent Gaussians Lemma 19, gives that w.h.p,

$$\max_{i \in S} c_i(0) \leq \max_{i \in [n]} c_i(0) \leq \sigma_{\text{init}} \sqrt{\frac{2 \log n}{n}}.$$

Finally combining everything yields

$$\frac{1}{\sqrt{\Phi}} \max_{i \in S} c_i(0) \leq \frac{\sigma_{\text{init}}}{4} \cdot \sigma_{\text{init}} \sqrt{\frac{2 \log n}{n}} = O(n^{-1/2} \log n)$$

w.h.p as desired.

Lemma 17 (Chi-square Tail Bound [24]) *If $X \sim \chi^2(k)$ then for all $t \in (0, 1)$,*

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

Lemma 18 (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*

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

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

Lemma 19 (Maximum of Gaussians) *Let $X_1, \ldots, X_n \sim_{iid} \mathcal{N}(0, \sigma^2)$. Then,*

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

A.3. Auxiliary Lemmas

Lemma 20 *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 21 *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. ■

Appendix B. Proofs for Section 2.4

B.1. Dynamics of Cyclic SGD

First let recall our setting. We assume that 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 $z_t = \langle \boldsymbol{w}_t, \boldsymbol{a}_1 \rangle$. From Eqs. (3, 4) it follows that the dynamics are given by

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

for $t\%2 = 0$ and

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

for $t\%2 = 1$. For convenience we let $F : \mathbb{R}^2 \to \mathbb{R}^2$ denote the function which gives the two-step update $(y_{t+2}, z_{t+2}) = F(y_t, z_t)$ for $t\%2 = 0$. We will make use of the following definitions

- Define the potential function $V(y,z) = z/y$. We will show that under the given initial conditions the potential is always decreasing after every two-steps.
- 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 two-steps.
- Define $\mathcal{A} = \{(y,z) \in (0,1)^2 : y \geq z > 0, y^2 + z^2 \leq 1 + \alpha/4\}$. 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 $F(y,z) \in \mathcal{A}$.

B.2. Proof of Theorem 4

We first begin with the observation that $y_t, z_t \in (0,1)$ for all t.

Lemma 22 *If $y_0, z_0 \in (0,1)$ and $\alpha \leq 1/4$ then $y_t, z_t \in (0,1)$ for all t.*

Proof We induct on t. If $t\%2 = 0$, then $y_{t+1} = y_t(1+\alpha(2-2y_t^2-z_t^2) \geq y_t(1-\alpha) > 0$ and similarly $z_{t+1} = z_t(1-\alpha y_t^2) \geq z_t(1-\alpha) > 0$. Therefore $y_{t+1}, z_{t+1} > 0$. Furthermore $z_{t+1} < z_t < 1$ and $y_{t+1} < y_t(1 + 2\alpha(1-y_t^2)) \leq 1$ by Lemma 30. The case for $t\%2 = 1$ is analogous. ■

Now we begin by with some observations about the behavior of the squared norm $N_t = y_t^2 + z_t^2$ in Lemma 24, 25, and 26. The first result Lemma 24 says that N_t is strictly increasing while $N_t < 1$ since $y_t > 0$ by Lemma 22. The second result Lemma 25 says that if at some point $N_t \geq 1$, then $N_{t'} \geq 1$ for all $t' \geq t$. Therefore since $N_0 < 1$ it follows that $N_t \geq N_0$ for all t. The last result Lemma 26 states that N_t is always at most $1 + \alpha/4$, so in fact $N_0 \leq N_t \leq 1 + \alpha/4$ for all t.

Now 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 $\mathcal{A}$ from Appendix B.1. In Proposition 27 we show that $\mathcal{A} \subseteq \mathcal{V}_-$. Then by Lemmas 22 and 26, along with 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 $V(y_{2t}, z_{2t}) \to V_\star \geq 0$.

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 $N_0 < 1$, by Lemmas 24, 25, and 26, we have that for all t, $0 < N_0 \leq N_{2t} \leq 1 + \alpha/4$. Since $V_\star \leq z_{2t}/y_{2t} \leq z_0/y_0 \leq 1$, the sequence $\{(y_{2t}, z_{2t})\} \subseteq \mathcal{K}_1 \subseteq \mathcal{V}_-$ where $\mathcal{K}_1 = \{(r\cos\theta, r\sin\theta) : \theta \in [\arctan(V_\star), \pi/4], r \in [N_0, 1 + \alpha/4]\}$. Since $\mathcal{K}_1$ is a compact set this is a contradiction by Proposition 23, therefore $V_\star = 0$.

Now we claim that $\liminf N_{2t} \geq 1$. If there exists t_0 such that $N_{t_0} \geq 1$, then by Lemma 25 we have $N_t \geq 1$ for all $t \geq t_0$, hence $\liminf N_{2t} \geq 1$. If however, $N_t < 1$ for all t, then by Lemma 24 N_{2t} is an increasing sequence, hence $\lim N_{2t} = \sup N_{2t}$. We claim that $\sup N_{2t} \geq 1$. Suppose for the sake of contradiction $N_\star := \sup N_{2t} < 1$. Then $\{(y_{2t}, z_{2t})\} \subseteq \mathcal{K}_2$ where $\mathcal{K}_2 = \{(r\cos\theta, r\sin\theta) : r \in [N_0, N_\star], \theta \in [0, \pi/4]\}$. By Lemma 24 it follows that for any $(y,z) \in \mathcal{K}_2$, $N(F(y,z)) - N(y,z) > 0$ therefore by Proposition 23 with $V = -N$ we get a contradiction.

Now we show that $\lim(y_{2t}, z_{2t}) = (1, 0)$. By Lemma 22

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

which implies that $\liminf y_{2t}^2 \geq \liminf N_{2t} - \lim(z_{2t}/y_{2t})^2 = 1$ and 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 29 which shows the gradient norm is continuous at the limit point, we see that $(y_t, z_t) \to (1, 0)$ as desired.

B.3. Auxiliary Results

Proposition 23 *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 $t = 0, 1 \ldots$ 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

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

clearly cannot hold since by compactness the left-hand side is finite and the right-hand side approaches negative infinity for large enough T. ■

Lemma 24 *Define $N_t = y_t^2 + z_t^2$. Assume that $N_t < 1$ and that $y_t, z_t \in [0, 1]$. Then $N_{t+1} \geq N_t$. The inequality is strict if $t\%2 = 0$ and $y_t > 0$ or $t\%2 = 1$ and $z_t > 0$.*

Proof Let $u_t = y_t^2$ and $v_t = z_t^2$. Assume $t\%2 = 0$. Then

$$\begin{aligned}
u_{t+1} - u_t &= \alpha u_t (2 - 2u_t - v_t)(2 + \alpha(2 - 2u_t - v_t)) \\
&= 2\alpha u_t (2 - 2u_t - v_t) + \alpha^2 u_t (2 - 2u_t - v_t)^2 \\
&= 2\alpha u_t (2 - 2u_t - v_t) + \alpha^2 u_t (2 - 2u_t - 2v_t + v_t)^2 \\
&\geq 2\alpha u_t (2 - 2u_t - v_t) + \alpha^2 u_t v_t^2 \\
v_{t+1} - v_t &= -\alpha v_t u_t (2 - \alpha u_t) \geq -2\alpha v_t u_t
\end{aligned}$$

Therefore

$$N_{t+1} - N_t = (u_{t+1} - u_t) + (v_{t+1} - v_t) \geq 2\alpha u_t (2 - 2u_t - 2v_t) + \alpha^2 u_t v_t^2 \geq 4\alpha u_t (1 - N_t)$$

from which the claim easily follows. The case for $t\%2 = 1$ follows by symmetry. ■

Lemma 25 *Assume $\alpha \leq 1/4$ and $y_t, z_t \in [0,1]$. If $y_t^2 + z_t^2 \geq 1$, then $y_{t+1}^2 + z_{t+1}^2 \geq 1$.*

Proof Let $N_t = u_t + v_t$. Observe that if $t\%2 = 0$ then

$$\begin{aligned} N_{t+1} = u_{t+1} + v_{t+1} &= u_t(1 + \alpha(2 - 2u_t - v_t))^2 + v_t(1 - \alpha u_t)^2 \\ &\geq u_t(1 + 2\alpha(2 - 2u_t - v_t)) + v_t(1 - 2\alpha u_t) \\ &= u_t + v_t + 2\alpha u_t(2 - 2u_t - 2v_t) \end{aligned}$$

The above inequality can be written as

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

Note that

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

if and only if

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

which is true since $N_t \leq 1$ and

$$\alpha \leq \frac{1}{4u_t} \leq 1/4$$

by assumption. The case for $t\%2$ follows by symmetry. ■

Lemma 26 *Assume $\alpha \leq 1/4$. If $y_0^2 + z_0^2 \leq 1 + \alpha/4$, then $y_t^2 + z_t^2 \leq 1 + \alpha/4$ for all t.*

Proof We prove this by induction. Let $N_t = y_t^2 + z_t^2$. Assume $N_t \leq 1 + \alpha/4$. Let $u_t = y_t^2$ and $v_t = z_t^2$. Assume $t\%2 = 0$. Then we have

$$N_{t+1} = u_t(1 + \alpha(2 - N_t - u_t))^2 + (N_t - u_t)(1 - \alpha u_t)^2.$$

Let

$$\begin{aligned} f(N; \alpha, u) &= u(1 + \alpha(2 - N - u))^2 + (N - u)(1 - \alpha u)^2 \\ &= u\alpha(2 - N)(\alpha(2 - N) + 2(1 - \alpha u)) + N(1 - \alpha u)^2 \end{aligned}$$

Note that $f''(N) \geq 0$ hence

$$\max_{N \in [0, 1+\alpha/4]} f(N) = \max(f(0), f(1 + \alpha/4)).$$

Note that

$$f(0) = u(1+\alpha(2-u))^2 - u(1-\alpha u)^2 = 4\alpha u(1+\alpha(1-u)) \leq \sup_{u \in [0,1]} 4\alpha u(1+\alpha(1-u)) = 4\alpha \leq 1.$$

Now fix $N = 1 + \alpha/4$ and note that be re-arranging

$$f(N) = \alpha^2(3N - 4)u^2 + \alpha u[\alpha(2 - N)^2 + 4(1 - N)] + N.$$

Note that

$$3N - 4 = 3 - 3\alpha/4 - 4 = -1 - 3\alpha/4 \leq 0$$

and

$$\alpha(2-N)^2+4(1-N)=\alpha(1-\alpha/4)^2-\alpha\leq 0$$

hence it is clear that $f(N)\leq N$. By symmetry (swapping v_t for u_t), an analogous result holds for $t\%2=1$. ■

Proposition 27 *The set* $\mathcal{A}=\{(y,z):y\geq z>0, y^2+z^2\leq 1+\alpha/4\}\subseteq\mathcal{V}_-$.

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

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

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

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

or equivalently

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

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

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

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

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

By Lemma 32

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

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

$$\frac{\mathrm{d}}{\mathrm{d}\theta}[f(y(\theta);r)-g(y(\theta);r)]=\frac{\mathrm{d}}{\mathrm{d}y}[f(y;r)-g(y;r)]\frac{\mathrm{d}y}{\mathrm{d}\theta}=\frac{\mathrm{d}}{\mathrm{d}y}[f(y;r)-g(y;r)](-r\sin\theta)\geq 0. \tag{18}$$

As $\cos(\pi/4)=\sin(\pi/4)=1/\sqrt{2}$, Lemma 28 states that if $r\leq\sqrt{1+\alpha/4}$ 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. (18) for $0 \le \psi \le \pi/4$

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

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

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

this proves the claim. ■

Lemma 28 *If* $0 < y^2 \le \frac{1}{2}(1+\alpha/4)$ *and* $\alpha \le 1/4$*, 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 $y^2 \le (1+\alpha/4)/2$. We have that

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

Therefore

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

Thus we have that

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

Substituting and factoring yields

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

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

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

Letting $b = 1 - \alpha$, it follows that $\alpha(1-y^2) = w - b$ 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 31. ■

Lemma 29 *Let t be even and $\alpha \leq 1/4$. Then if $\max(1 - y_t, z_t) < \varepsilon < 1$ then $\max(1 - y_{t+1}, z_{t+1}) < 2\varepsilon$.*

Proof Just check the size of gradients

- $\alpha y_t z_t^2 \leq \alpha \varepsilon^2 < \varepsilon$.
- $\alpha z_t (2 - 2z_t^2 - y_t^2) \leq 2\alpha\varepsilon < \varepsilon$.

■

B.4. Technical Lemmas

Lemma 30 *For $\eta \in (0, 1/2]$, $\sup_{x \in [0,1]} x(1 + \eta(1 - x^2)) = 1$.*

Proof Let $f(x) = x(1 + \eta(1 - x^2))$, then $f'(x) = 1 + \eta - 3\eta x^2$. Note that $f'(x) \geq 0$ iff

$$x^2 \leq \frac{1 + \eta}{3\eta}$$

and that since $\eta \leq 1/2$ implies

$$\frac{1 + \eta}{3\eta} \geq 1$$

we see that $f'(x) \geq 0$ if $x^2 \leq 1$, therefore $\sup_{x \in [0,1]} f(x) = f(1) = 1$. ■

Lemma 31 *Assume $\alpha \leq 1/4$ and $2y^2 \leq 1 + \alpha/4$. Let $w = 1 - \alpha y^2$ and $b = 1 - \alpha$. 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 $y^2 \leq (1 + \alpha/4)/2$,

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

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

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

which is true since by assumption $\alpha \leq 1/4 \leq 2/7$. 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 32 *Assume $r \leq 1 + \alpha/4$ is a constant. Define the following functions of $y \in [0, 1]$.*

$$f(y; r) = (1 - \alpha y^2)(1 + \alpha[2 - 2(r - y^2)(1 - \alpha y^2)^2 - y^2(1 + \alpha(2 - r - y^2))^2])$$
$$g(y; r) = (1 - \alpha(r - y^2)(1 - \alpha y)^2)(1 + \alpha(2 - r - y^2)).$$

Then the following is true

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

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

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

Using Mathematica we have that

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

where

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

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

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

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

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

Therefore we have shown that

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

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

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

■

Appendix C. Proofs for Section 2.5

C.1. Proof of Theorem 7

Proof It is clear that when considering the minimum of the loss objective we can restrict our consideration to $\boldsymbol{w} \in \mathrm{span}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m)$. Let $\boldsymbol{w} = \sum_{i=1}^{m} c_i \boldsymbol{a}_i$. 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

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

Therefore to find a minimizer it suffices to minimize the quantity

$$-2\sum_{i=1}^{m} c_i\phi(c_i) + \sum_{i=1}^{m} c_i^2\sum_{i=1}^{m}\phi(c_i)^2. \tag{19}$$

If we define

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

then Eq. (19) 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. ■

C.2. Hessians of Critical Points

Let us compute the Hessian of the loss function at a point $\boldsymbol{w}$.

Lemma 33 *The Hessian of the loss $\mathcal{L}$ at $\boldsymbol{w}$ is given by*

$$\nabla^2\mathcal{L}(\boldsymbol{w}) = \frac{1}{m}\sum_{\ell\in[m]}\phi'(\langle\boldsymbol{w}, \boldsymbol{a}_\ell\rangle)\left[(\|\boldsymbol{w}\|^2 - 2)\cdot\boldsymbol{a}_\ell\boldsymbol{a}_\ell^\top + 2\langle\boldsymbol{w}, \boldsymbol{a}_\ell\rangle(\boldsymbol{a}_\ell\boldsymbol{w}^\top + \boldsymbol{w}\boldsymbol{a}_\ell^\top) + \langle\boldsymbol{w}, \boldsymbol{a}_\ell\rangle^2\cdot\mathbf{I}\right].$$

Proof For shorthand we will define $\mathbb{E}_{\boldsymbol{x}}$ as the expectation over $\boldsymbol{x} \sim \mathsf{Unif}(\mathcal{D})$.

$$\begin{aligned}
\nabla^2\mathcal{L}(\boldsymbol{w}) &= \mathbb{E}_{\boldsymbol{x}}\phi'(\langle\boldsymbol{w}, \boldsymbol{x}\rangle)\left[(\|\boldsymbol{w}\|^2 - 2)\cdot\boldsymbol{x}\boldsymbol{x}^\top + 2\langle\boldsymbol{w}, \boldsymbol{x}\rangle(\boldsymbol{x}\boldsymbol{w}^\top + \boldsymbol{w}\boldsymbol{x}^\top) + \langle\boldsymbol{w}, \boldsymbol{x}\rangle^2\cdot\mathbf{I}\right] \\
&= \frac{1}{m}\sum_{\ell\in[m]}\phi'(\langle\boldsymbol{w}, \boldsymbol{a}_\ell\rangle)\left[(\|\boldsymbol{w}\|^2 - 2)\cdot\boldsymbol{a}_\ell\boldsymbol{a}_\ell^\top + 2\langle\boldsymbol{w}, \boldsymbol{a}_\ell\rangle(\boldsymbol{a}_\ell\boldsymbol{w}^\top + \boldsymbol{w}\boldsymbol{a}_\ell^\top) + \langle\boldsymbol{w}, \boldsymbol{a}_\ell\rangle^2\cdot\mathbf{I}\right].
\end{aligned}$$

■

Proof [Proof of Proposition 8] Let $\boldsymbol{w} := \boldsymbol{w}_\star^{\mathrm{GD}} = \sum_{i\in S} \frac{\langle \boldsymbol{w}_0, \boldsymbol{a}_i\rangle}{\sqrt{\Phi}} \boldsymbol{a}_i$. Plugging into Lemma 33

$$\begin{aligned}
\boldsymbol{H}_{\mathrm{GD}} &= \frac{1}{m}\sum_{\ell\in S}\left[-\boldsymbol{a}_\ell\boldsymbol{a}_\ell^\top + \frac{2\langle \boldsymbol{w}_0, \boldsymbol{a}_\ell\rangle}{\sqrt{\Phi}}\left(\sum_{i\in S}\frac{\langle \boldsymbol{w}_0, \boldsymbol{a}_i\rangle}{\sqrt{\Phi}}(\boldsymbol{a}_i\boldsymbol{a}_\ell^\top + \boldsymbol{a}_\ell\boldsymbol{a}_i^\top)\right) + \frac{\langle \boldsymbol{w}_0, \boldsymbol{a}_\ell\rangle^2}{\Phi}\cdot\mathbf{I}\right] \\
&= \frac{1}{m}\sum_{\ell\in S}\left[-\boldsymbol{a}_\ell\boldsymbol{a}_\ell^\top + \frac{2\langle \boldsymbol{w}_0, \boldsymbol{a}_\ell\rangle}{\Phi}\sum_{i\in S}\langle \boldsymbol{w}_0, \boldsymbol{a}_i\rangle(\boldsymbol{a}_i\boldsymbol{a}_\ell^\top + \boldsymbol{a}_\ell\boldsymbol{a}_i^\top) + \frac{\langle \boldsymbol{w}_0, \boldsymbol{a}_\ell\rangle^2}{\Phi}\cdot\mathbf{I}\right] \\
&= -\frac{1}{m}\sum_{\ell\in S}\boldsymbol{a}_\ell\boldsymbol{a}_\ell^\top + \frac{2}{m}\frac{1}{\Phi}\sum_{\ell\in S}\sum_{i\in S}\langle \boldsymbol{w}_0, \boldsymbol{a}_\ell\rangle\langle \boldsymbol{w}_0, \boldsymbol{a}_i\rangle(\boldsymbol{a}_i\boldsymbol{a}_\ell^\top + \boldsymbol{a}_\ell\boldsymbol{a}_i^\top) + \frac{1}{m}\cdot\mathbf{I} \\
&= -\frac{1}{m}\sum_{\ell\in S}\boldsymbol{a}_\ell\boldsymbol{a}_\ell^\top + \frac{4}{m}\frac{1}{\Phi}\sum_{\ell\in S}\sum_{i\in S}\langle \boldsymbol{w}_0, \boldsymbol{a}_\ell\rangle\langle \boldsymbol{w}_0, \boldsymbol{a}_i\rangle\boldsymbol{a}_i\boldsymbol{a}_\ell^\top + \frac{1}{m}\cdot\mathbf{I}.
\end{aligned}$$

Similarly if $\boldsymbol{w} := \boldsymbol{w}_\star^{\mathrm{SGD}} = \boldsymbol{a}_k$ for some $k \in S$ then

$$\boldsymbol{H}_{\mathrm{SGD}} = \frac{1}{m}\left[-\boldsymbol{a}_k\boldsymbol{a}_k^\top + 4\boldsymbol{a}_k\boldsymbol{a}_k^\top + \mathbf{I}\right] = \frac{3\boldsymbol{a}_k\boldsymbol{a}_k^\top + \mathbf{I}}{m}.$$

■

Now we compute the eigenspectra of $\boldsymbol{H}_{\mathrm{GD}}$ and $\boldsymbol{H}_{\mathrm{SGD}}$. The eigenspectrum of $\boldsymbol{H}_{\mathrm{SGD}}$ is trivial, hence we only prove the result for $\boldsymbol{H}_{\mathrm{GD}}$.

Proof [Proof of Lemma 9] For convenience let $\boldsymbol{H} = \boldsymbol{H}_{\mathrm{GD}}$ and define

$$c_\ell := \frac{\langle \boldsymbol{w}_0, \boldsymbol{a}_\ell\rangle}{\sqrt{\Phi}}, \quad \ell \in S.$$

Note that by the definition of Φ in Eq. (6), we have $\sum_{\ell\in S} c_\ell^2 = 1$. Consider a unit vector $\boldsymbol{v} \in \mathrm{span}(\boldsymbol{a}_\ell : \ell \in S)$ i.e. $\boldsymbol{v} = \sum_{\ell\in S} b_\ell\boldsymbol{a}_\ell$ with $\sum_{\ell\in S} b_\ell^2 = 1$. Then

$$\begin{aligned}
\boldsymbol{H}\boldsymbol{v} &= \left(\frac{4}{m}\sum_{\ell\in S}\sum_{i\in S} c_\ell c_i \boldsymbol{a}_i\boldsymbol{a}_\ell^\top\right)\boldsymbol{v} \\
&= \frac{4}{m}\sum_{\ell\in S}\sum_{i\in S}\sum_{j\in S} c_\ell c_i b_j \boldsymbol{a}_i\boldsymbol{a}_\ell^\top\boldsymbol{a}_j \\
&= \frac{4}{m}\sum_{\ell\in S} c_\ell b_\ell \sum_{i\in S} c_i\boldsymbol{a}_i.
\end{aligned}$$

Therefore if $b_\ell = c_\ell$ then $\boldsymbol{H}\boldsymbol{v} = (4/m)\boldsymbol{v}$ and if $\sum_{\ell\in S} b_\ell c_\ell = 0$ then $\boldsymbol{H}\boldsymbol{v} = \mathbf{0}$. This gives $|S|$ orthogonal eigenvectors. Note that if $\boldsymbol{v}$ is orthogonal to $\mathrm{span}(\boldsymbol{a}_\ell : \ell \in S)$ then $\boldsymbol{v}$ is an eigenvector with eigenvalue of $1/m$. ■