ON GRADIENT DESCENT CONVERGENCE BEYOND THE EDGE OF STABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient Descent (GD) is a powerful workhorse of modern machine learning thanks to its scalability and efficiency in high-dimensional spaces. Its ability to find local minimisers is only guaranteed for losses with Lipschitz gradients, where it can be seen as a 'bona-fide' discretisation of an underlying gradient flow. Yet, many ML setups involving overparametrised models do not fall into this problem class, which has motivated research beyond the so-called "Edge of Stability", where the step-size crosses the admissibility threshold inversely proportional to the Lipschitz constant above. Perhaps surprisingly, GD has been empirically observed to still converge regardless of local instability.

In this work, we study a local condition for such an unstable convergence around a local minima in a low dimensional setting. We then leverage these insights to establish global convergence of a two-layer single-neuron ReLU student network aligning with the teacher neuron in a large learning rate beyond the Edge of Stability under population loss. Meanwhile, while the difference of norms of the two layers is preserved by gradient flow, we show that GD above the edge of stability induces a balancing effect, leading to the same norms across the layers.

1 INTRODUCTION

Given a differentiable objective function $f(\theta)$, where $\theta \in \mathbb{R}^d$ is a high-dimensional parameter vector, the most basic and widely used optimization method is gradient descent (GD), defined as

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_\theta f(\theta^{(t)}), \tag{1}$$

where η is the learning rate. For all its widespread application across many different ML setups, a basic question remains: what are the convergence guarantees (even to a local minimiser) under typical objective functions, and how they depend on the (only) hyperaparameter η?

In recent decades, neural networks have been a popular parameterized model for its fascinating generalization ability. But it still remains an open problem to understand its loss landscape and training behavior under GD. The sensible regime to start the analysis is $\eta \to 0$, where GD inherits the local convergence properties of the Gradient Flow ODE via standard arguments from numerical integration. However, in the early phase of training, a large learning rate has been observed to result in better generalization (LeCun et al., 2012; Bjorck et al., 2018; Jiang et al., 2019; Jastrzebski et al., 2021), where the extent of "large" is measured by comparing the learning rate η and the curvature of loss landscape λ, *i.e.*, the largest eigenvalue of the training loss Hessian with respect to learnable parameters. Although one requires $\lambda < 2/\eta$ to guarantee the (local) convergence of GD (Bottou et al., 2018), it has been empirically shown that, for a fixed η, the curvature λ increases along the training trajectory, which brings $\lambda \geq 2/\eta$ (Cohen et al., 2020). After that, a surprising phenomena is that λ *stably* hovers above $2/\eta$ and the neural network still achieves a decreasing training loss from a broader view of time — the so-called "Edge of Stability". We would like to understand and analyse the conditions of such convergence with a large learning rate under a variety models that capture such observed empirical behavior.

The straightforward starting point for the local convergence analysis is via Taylor approximations of the loss function. However, in quadratic Taylor expansion, gradient descent diverges once $\lambda > 2/\eta$ (Cohen et al., 2020), which requires a higher order approximation. By considering a 1-D

function with local minima of curvature λ, we show that it is possible to stably oscillate around the minima with η slightly above the threshold $2/\lambda$, provided its high order derivative satisfies mild conditions. A typical example of such functions is $f(x) = \frac{1}{4}(x^2-\mu)^2$ with $\mu > 0$, and the third-order derivative around the minima is $f^{(3)}(\sqrt{\mu}) = 6x \neq 0$. Furthermore, we prove that it converges to an orbit of period 2 from a more global view rather than the analysis of high-order local approximation.

A natural follow-up question is what implications such a 1-D function brings to our understanding of neural network training, which is in high dimension. Intuitively, we would like to emphasize that a non-zero third-order derivative implies asymmetricity in the loss landscape, and He et al. (2019) empirically proposes that prevalently there exist asymmetric valleys in the landscape. We extend our analysis of $f(x) = \frac{1}{4}(x^2 - \mu)^2$ to a 2-D function $g(x, y) = \frac{1}{2}(xy - \mu)^2$. For this function, if GD is equipped with $\eta > 2/\lambda$, then x and y converge to a balanced setting $x = y$, where $\lambda \triangleq \lambda_{\max}(\nabla^2 g(\sqrt{\mu}, \sqrt{\mu}))$. On one hand, with balanced x and y, the trajectories of $f(x)$ and $g(x, y)$ are aligned. On the other hand, the loss landscape is flatter when parameters are balanced (Elkabetz & Cohen, 2021). However, different from a previous result (Du et al., 2018) that gradient flow preserves $x - y$ along the training — and hence that a balanced initialization guarantees a balanced trajectory, our analysis reveals that *inbalanced* initialization ends in balance with GD beyond the edge of stability. This result not only compresses the problem into low dimension but also sheds light on the connection of GD with landscape curvature, complementing existing insights (Jastrzebski et al., 2019; Arora et al., 2022; Cohen et al., 2020).

We finally leverage the 2-D analysis of $g(x, y)$ to a two-layer single-neuron ReLU network, where the task is to learn a teacher neuron with data on a uniform high-dimensional sphere. We show a convergence result under population loss with GD beyond the edge of stability, where the direction of the teacher neuron can be learnt and the norms of two-layer weights stably oscillate as in the 2-D case.

In summary, in the setting of gradient descent of a learning rate $\eta > 2/\lambda$, we prove that:

- In a general 1-D function, there exists a condition on the high-order derivatives so that gradient descent stably oscillates around the local minima.
- In a specific 1-D function in the form of $f(x) = (x^2 - \mu)^2$, the parameter converges to an orbit of period 2 under mildly global initialization.
- In a specific 2-D function in the form of $g(x, y) = (xy - \mu)^2$, the two parameters converge to balance under mildly inbalanced initialization.
- A two-layer single-neuron ReLU student network is able to find the correct direction of the teacher neuron. Then it shares the properties with the 2-D case.

We also run numerical experiments to support the above theorems. Note that the learning rate is also upper-bounded in the theorems, but our contribution is that its lower bound is beyond the common wisdom.

2 Related Work

Implicit regularization. Due to its theoretical closeness to gradient descent with a small learning rate, gradient flow is a common setting to study the training behavior of neural networks. Barrett & Dherin (2020) suggests that gradient descent is closer to gradient flow with an additional term regularizing the norm of gradients. Through analysing the numerical error of Euler's method, Elkabetz & Cohen (2021) provides theoretical guarantees of a small gap depending on the convexity along the training trajectory. Neither of them fits in the case of our interest, because it is hard to track the parametric gap when $\eta > 1/\lambda$. For instance, in a quadratic function, the trajectory jumps between the two sides once $\eta > 1/\lambda$. Damian et al. (2021) shows that SGD with label noise is implicitly subjected to a regularizer penalizing sharp minimizers but the learning rate is constraint strictly below the edge of stability threshold.

Balancing effect. Du et al. (2018) proves that gradient flow automatically preserves the norms' differences between different layers of a deep homogeneous network. (Ye & Du, 2021) shows that gradient descent on matrix factorization with a constant small learning rate still enjoys the

auto-balancing property. Also in matrix factorization, Wang et al. (2021) proves that gradient descent with a relatively large learning rate leads to a solution with a more balanced (perhaps not perfectly balanced) solution while the initialization can be in-balanced. In a similar spirit, we extend their finding to a larger learning rate, with which the perfect balance may be achieved in our setting. We estimate our learning rate is at least $3\times$ theirs (Wang et al., 2021). Note that the implication of balancing effect is to get close to a flatter solution in the global minimum manifold, which may help improve generalization in some common arguments in the community.

Edge of stability. Cohen et al. (2020) observes a two-stage process in gradient descent, where the first is loss curvature grows until the sharpness touches the bound $2/\eta$, and the second is the curvature hovers around the bound and training loss still decreases in a macro view regardless of local instability. Gilmer et al. (2021) reports similar observations in stochastic gradient descent and conducts comprehensive experiments of loss sharpness on learning rates, architecture choices and initialization. Lewkowycz et al. (2020) argues that gradient descent would "catapult" into a flatter region if loss landscape around initialization is too sharp.

Some concurrent works (Ahn et al., 2022; Ma et al., 2022; Arora et al., 2022) are also theoretically investigating the edge of stability. Ahn et al. (2022) suggests that unstable convergence happens when the loss landscape of neural networks forms a local forward-invariant set near the minima due to some ingredients, such as $\tanh$ as the nonlinear activation. Ma et al. (2022) empirically observes a multi-scale structure of loss landscape and, with it as an assumption, shows that gradient descent with different learning rates may stay in different levels. Arora et al. (2022) shows the training provably enters the edge of stability with modified gradient descent or modified loss, and then its associated flow goes to flat regions.

Learning a single neuron. Yehudai & Ohad (2020) studies necessary conditions on both the distribution and activation functions to guarantee a one-layer single student neuron aligning with the teacher neuron under gradient descent, SGD and gradient flow. Vardi et al. (2021) extends the investigation into a neuron with a bias term. Vardi & Shamir (2021) empirically studies the training dynamics of a two-layer single neuron, focusing on its implicit bias. In this work, we present a convergence analysis of a two-layer single-neuron ReLU network trained with population loss in a large learning rate beyond the edge of stability.

3 Problem Setup

We consider a differentiable objective function $f(\theta)$ with $\theta \in \mathbb{R}^d$, and the GD algorithm from (1).

Definition 1. *A differentiable function f is L-gradient Lipschitz if*

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)\| \leq L \|\theta_1 - \theta_2\|, \quad \forall\, \theta_1, \theta_2. \tag{2}$$

The above definition is equivalent to saying that the spectral norm of the Hessian is bounded by L, or the *local curvature* at each point is bounded by L. Then η needs to be bounded by $1/L$ in GD so that it is guaranteed to visit an approximate first-order stationary point (Nesterov, 1998). The perturbed GD requires $\eta = 1/L$ to visit an approximate second-order stationary point (Jin et al., 2021), and stochastic variants share similar assumptions (Ghadimi & Lan, 2013; Jin et al., 2021).

However, in practice, such an assumption may be violated, or even impossible to satisfy when $\|\nabla^2 f\|$ is not uniformly bounded. Cohen et al. (2020) observes that, with learning rate η fixed, the largest eigenvalue λ_1 of the loss Hessian of a neural network is below $2/\eta$ at initialization, but it grows above the threshold along training. Such a phenomena is more obvious when the network is deeper or narrower. This reveals the non-smooth nature of the loss landscape of neural networks.

Furthermore, another observation from Cohen et al. (2020) is that once $\lambda_1 \geq 2/\eta$, the training loss starts to perturb sharply. This is not surprising because GD would diverge in a quadratic function with such a large curvature. However, despite of local instability, the training loss still decreases in a longer range of steps, during which the curvature L stays around $2/\eta$. A further phenomena is that, when GD is at the edge of stability, if the learning rate suddenly changes to a smaller value $\eta_s < \eta$, then the local curvature quickly grows to $2/\eta_s$ — indicating the ability to 'manipulate' the local curvature by adjusting the learning rate.

Separated from GD, the local curvature itself enjoys a lot of attention. Due to the nature of over-parameterization in modern neural networks, the global minimizers of the objective f form a manifold of solutions. There have been active directions to understand the *implicit bias* of GD methods, namely where do they converge to in the manifold, and why some points in the manifold are more preferable than others. For the former question, it is believed that (stochastic) GD prefers flatter minima (Barrett & Dherin, 2020; Smith et al., 2021; Damian et al., 2021; Ma & Ying, 2021). For the latter, flatter minima brings better generalization (Hochreiter & Schmidhuber, 1997; Li et al., 2018; Keskar et al., 2016; Ma & Ying, 2021; Ding et al., 2022). It would be meaningful if flatter minima could be obtained via GD with a large learning rate.

More specifically, it has been shown that the eigenvalues of the hessian of a deep homogeneous network could be manipulated to infinity via rescaling the weights of each layer (Elkabetz & Cohen, 2021). Fortunately, gradient flow preserves the difference of norms across layers along the training (Du et al., 2018). As a result, a balanced initialization induces balanced convergence, while GD would break this balancing effect due to finite learning rate. However, recently it has been observed that GD with large learning rates enjoys a balancing effect (Wang et al., 2021), where it converges to a (not perfect) balanced result despite of inbalanced initialization.

Motivated by the connections of optimization, loss landscape and generalization, we would like to understand the training behavior and the balancing effect of gradient descent with a large learning rate.

4 Stable oscillation on 1-D and 2-D functions

In this section we study the phenomenon of stable oscillation on 1-D and 2-D functions. Starting from a condition on general 1-D functions, we look into several specific 1-D functions to verify our arguments. Then, focusing on a function in the form of $f(x) = (x^2 - \mu)^2$, we present the convergence analysis as a foundation for the following discussions. Furthermore, on a 2-D function, we show a balancing effect to make a connection to the 1-D analysis.

4.1 On 1-D functions

General 1-D functions. Consider a 1-D function $f(x)$ with a learnable parameter $x \in \mathbb{R}$. The parameter updates following GD with the learning rate η as

$$x^{(t+1)} := x^{(t)} - \eta f'(x^{(t)}). \tag{3}$$

Assuming f is differentiable and all derivatives are bounded, the function value in the next step can be approximated by

$$f(x^{(t+1)}) = f(x^{(t)}) - \eta[f'(x^{(t)})]^2\Big(1 - \frac{\eta}{2}f''(x^{(t)})\Big) + o((x^{(t+1)} - x^{(t)})^2). \tag{4}$$

If $\eta < 2/f''(x^{(t)})$, it reveals the function monotonely decreases for each step of GD, ignoring higher terms. Such an assumption would guarantee the convergence to a global minimum in a convex function. However, our interest is what happens if $\eta > 2/f''(x)$. For instance, if f is a quadratic function, the second-order derivative f'' is constant. As a result, once $\eta > 2/f''$, GD diverges except being initialized at the optimum. However, when trained with a large learning rate $\eta > 2/f''(\bar{x})$, there is still some hope for a function to stay around a local minima $\bar{x}$, as stated in the following theorem.

Theorem 1. *Consider a 1-D differentiable function $f(x)$ around a local minima $\bar{x}$. Assume 1) $f^{(3)}(\bar{x}) \neq 0$, 2) $\frac{f^{(3)}(\bar{x})}{f''(\bar{x})} = \mathcal{O}(1)$, 3) $3[f^{(3)}]^2 - f''f^{(4)} > 0$ at $x = \bar{x}$, and 4) all higher order derivatives are bounded. For a start point near $\bar{x}$ as $x_0 = x - \epsilon$ such that $\epsilon \cdot f^{(3)}(\bar{x}) > 0$ and ϵ is sufficiently small, then there exists a learning rate η with which GD starting from x_0 will back to it in two steps, and*

$$\frac{2}{f''(\bar{x})} < \eta < \frac{2}{f''(\bar{x}) - \epsilon \cdot f^{(3)}(\bar{x})}.$$

The details of proof are presented in the Appendix A. As stated in the Theorem 1, we provide a sufficient condition that allows GD to stably oscillate around a local minima. But still we cannot tell

whether or not some functions allow it with $f^{(3)}(\bar{x}) = 0$. For instance, a quadratic function does not satisfy this condition since $f^{(3)} \equiv 0$ and it diverges when GD is beyond the edge of stability. For $f(x) = \sin(x)$ around $\bar{x} = -\frac{\pi}{2}$ where $f^{(3)}(\bar{x}) = 0$, it turns out the sine function allows stable oscillation. Therefore, we extend the argument in Theorem 1 to a higher order case in Lemma 1.

Lemma 1. *Consider a 1-D differentiable function $f(x)$ around a local minima $\bar{x}$. Assume 1) the lowest order non-zero derivative of f (except the f'') at $\bar{x}$ is $f^{(k)}(\bar{x})$, 2) $\frac{f^{(k)}(\bar{x})}{f''(\bar{x})} = \mathcal{O}(1)$, 3) all higher order derivatives are bounded, and 4) $k \geq 4$. Then, from a starting point near $\bar{x}$ as $x_0 = \bar{x} - \epsilon$ with $\epsilon > 0$ sufficiently small, and*

1. *if k is odd and $f^{(k+1)}(\bar{x}) < 0$, then there exists $\eta \in \left(\frac{2}{f''}, \frac{2}{f'' - f^{(k)}\epsilon^{k-2}}\right)$,*
2. *if k is even and $f^{(k)}(\bar{x}) < 0$, then there exists $\eta \in \left(\frac{2}{f''}, \frac{2}{f'' + f^{(k)}\epsilon^{k-2}}\right)$,*

such that: GD can stably oscillate in a sufficiently small neighborhood of $\bar{x}$ with η.

With Lemma 1, we can verify the sine function to allow stable oscillation as in Corollary 1, because its lowest order of nonzero derivative (except f'') at the local minima is $f^{(4)}(\bar{x}) < 0$. Meanwhile, Theorem 1 provides a guarantee that L_2 loss on any function g provably allows stable oscillation once g satisfies some mild conditions, as stated below.

Lemma 2. *Consider a 1-D function $g(x)$ with the a target value y. Then define the loss function f as $f(x) = (g(x) - y)^2$. Assuming 1) g' is not zero when $g(\bar{x}) = y$, 2) all higher order derivatives are well bounded, then it satisfies the condition in Theorem 1 or Lemma 1 to allow stable oscillation around $\bar{x}$. More precisely, it holds if*

$$g'(\bar{x})g^{(3)}(\bar{x}) < 6[g''(\bar{x})]^2.$$

Proof details of the above lemmas are presented in the Appendix B and C. This design of L_2 loss is to fit the target with a generic function, which is quite common in practice. The most significant condition is that the target value shall not be too special. For instance, any value $y \in (-1, 1)$ works for $g(x) = \sin(x)$ or $g(x) = \tanh(x)$. The proof details for these settings of $g(x)$ are provided as Corollaries 2, 3 and 4 at the end of Appendix C. Next we are going to present a careful analysis on $g(x) = x^2$.

A special 1-D function. Consider $f(x) = \frac{1}{4}(x^2 - \mu)^2$ with $\mu > 0$, $f^{(3)}(\sqrt{\mu}) = 6\sqrt{\mu}$, $f''(\sqrt{\mu}) = 2\mu$. Note that this function is more special to us because it can be viewed as a *symmetric scalar factorization* problem subjected to L_2 loss. Later we will develop it by two levels to gain insights for asymmetric initialization and two-layer single-neuron networks. Before that, we would like to show where it converges to when $\eta > \frac{2}{f''(\sqrt{\mu})}$ as follows.

Theorem 2. *For $f(x) = \frac{1}{4}(x^2 - \mu)^2$, consider GD with $\eta = K \cdot \frac{1}{\mu}$ where $1 < K \leq 1.121$. The process starts from any point $0 < x < \sqrt{\mu}$. Then it converges to an orbit of period 2, except for a measure-zero initialization where it converges to $\sqrt{\mu}$.*

The details of proof are presented in the Appendix D. As shown above, Theorem 1 and Theorem 2 stand in two different levels: Theorem 1 restricts the discussion in a local view because of Taylor approximation, while Theorem 2 starts from local convergence and then generalizes it into a global view. However, Theorem 1 builds a foundation for Theorem 2 because the latter would degenerate to the former when K is extremely close to 1.

A natural following question is what implications Theorem 2 brings, because 1-D is far from the practice of neural networks that contain multi-layer structures, nonlinearity and high dimension. We would like to incorporate multi layers first, as discussed below.

4.2 On a 2-D function

Similar to $f(x) = \frac{1}{4}(x^2 - \mu)^2$, consider a 2-D function $f(x, y) = \frac{1}{2}(xy - \mu)^2$. Apparently, if x and y initialize as the same, then $(x^{(t)}, y^{(t)})$ would always align with the 1-D case from the same initialization. Therefore, it is significant to analyze this problem under different initialization for x

and y, which we would call "in-balanced" initialization. Meanwhile, another giant difference is that all the global minima in 2-D case form a manifold $\{(x, y)|xy = \mu\}$ while the 1-D case only has two points of global minima. It would be great if we could understand which points in the global minima manifold, or in the whole parameter space, are preferable by GD.

Note that reweighting the two parameters would manipulate the curvature to infinity as in (Elkabetz & Cohen, 2021), so the inbalance strongly affects the local curvature. Viewing $f(x)$ as a symmetric scalar factorization problem, we treat $f(x, y)$ as asymmetric scalar factorization. The update rule of GD is

$$x^{(t+1)} := x^{(t)} - \eta(x^{(t)}y^{(t)} - \mu)y^{(t)}, \quad y^{(t+1)} := y^{(t)} - \eta(x^{(t)}y^{(t)} - \mu)x^{(t)}. \tag{5}$$

Consider the Hessian as

$$H \triangleq \begin{bmatrix} \partial_x^2 f & \partial_y \partial_x f \\ \partial_x \partial_y f & \partial_y^2 f \end{bmatrix} = \begin{bmatrix} y^2 & 2xy - \mu \\ 2xy - \mu & x^2 \end{bmatrix}. \tag{6}$$

When $xy = \mu$, the eigenvalues of H are $\lambda_1 = x^2 + y^2, \lambda_2 = 0$. Note that $\lambda_1 = (x - y)^2 + 2\mu$. Hence, in the global minima manifold, the local curvature of each point is larger if its two parameters are more inbalanced. Among all these points, the smallest curvature appears to be $\lambda_1 = 2\mu$ when $x = y = \sqrt{\mu}$. In other words, if the learning rate $\eta > 2/2\mu$, all points in the manifold would be too sharp for GD to converge. We would like to investigate the behavior of GD in this case. It turns out the two parameters are driven to a perfect balance although they initialized differently, as follows.

Theorem 3. *For $f(x, y) = \frac{1}{2}(xy - \mu)^2$, consider GD with learning rate $\eta = K \cdot \frac{1}{\mu}$. Assume both x and y are always positive during the whole process $\{x_i, y_i\}_{i\geq 0}$. In this process, denote a series of all points with $xy > \mu$ as $\mathcal{P} = \{(x_i, y_i)|x_i y_i > \mu\}$. Then $|x - y|$ decays to 0 in $\mathcal{P}$, for any $1 < K < 1.5$.*

Proof sketch The details of proof are presented in the Appendix E. Start from a point $(x^{(t)}, y^{(t)})$ where $x^{(t)}y^{(t)} > \mu$. Because $y^{(t+1)} - x^{(t+1)} = (y^{(t)} - x^{(t)})(1 + \eta(x^{(t)}y^{(t)} - \mu))$, it suffices to show

$$\left| \frac{y^{(t+2)} - x^{(t+2)}}{y^{(t)} - x^{(t)}} \right| = |(1 + \eta(x^{(t)}y^{(t)} - \mu))(1 + \eta(x^{(t+1)}y^{(t+1)} - \mu))| < 1. \tag{7}$$

Since $1 + \eta(x^{(t)}y^{(t)} - \mu) > 1$, the analysis of $1 + \eta(x^{(t+1)}y^{(t+1)} - \mu)$ is divided into three cases considering the coupling of $(x^{(t)}, y^{(t)}), (x^{(t+1)}, y^{(t+1)})$. □

Remark. Actually, for a larger $K \geq 1.5$, it is possible for GD to converge to an inbalanced orbit. For instance, Figure 15 in (Wang et al., 2021) shows inbalanced orbits for $f(x) = \frac{1}{2}(xy - 1)^2$ with $K = 1.9$.

Combining with the fact that the probability of GD converging to a stationary point that has sharpness beyond the edge of stability is zero (Ahn et al., 2022), Theorem 3 reveals x and y would converge to a perfect balance. Note that this balancing effect is different from that of gradient flow (Du et al., 2018), where the latter states that gradient flow preserves the difference of norms of different layers along training. As a result, in gradient flow, inbalanced initialization induces inbalanced convergence, while in our case inbalanced-initialized weights converge to a perfect balance. Furthermore, Theorem 3 shows an effect that the two parameters are squeezed to a single variable, which re-directs to our 1-D analysis in Theorem 2. Therefore, actually both cases converge to the same orbit when $1 < K \leq 1.121$. Numerical results are presented in Figure 2.

However, one thing to notice is that the inbalance at initialization needs to be bounded in Theorem 3 because both x and y are assumed to stay positive along the training. More precisely, we have

$$x^{(t+1)}y^{(t+1)} = x^{(t)}y^{(t)}(1 - \eta(x^{(t)}y^{(t)} - \mu))^2 - \eta(x^{(t)}y^{(t)} - \mu)(x^{(t)} - y^{(t)})^2, \tag{8}$$

and then $x^{(t+1)}y^{(t+1)} < 0$ when $|x^{(t)} - y^{(t)}|$ is large with $x^{(t)}y^{(t)} > \mu$ fixed. Therefore, we provide a condition to guarantee both x, y positive as follows, with details presented in the Appendix F.

Lemma 3. *In the setting of Theorem 3, denote the initialization as $m = \frac{|y_0 - x_0|}{\sqrt{\mu}}$ and $x_0 y_0 > \mu$. Then, during the whole process, both x and y will always stay positive, denoting $p = \frac{4}{\left(m + \sqrt{m^2 + 4}\right)^2}$ and $q = (1 + p)^2$, if*

$$\max\left\{\eta(x_0 y_0 - \mu), \frac{4}{27}(1 + K)^3 + \left(\frac{2}{3}K^2 - \frac{1}{3}K + \frac{qK^2}{2(K + 1)}m^2\right)qm^2 - K\right\} < p.$$

5 ON A TWO-LAYER SINGLE-NEURON HOMOGENEOUS NETWORK

We denote a two-layer single-neuron network as $f(x;\theta) = v \cdot \sigma(w^\top x)$ where $v \in \mathbb{R}, w \in \mathbb{R}^d$, the set of trained parameters $\theta = (v, w^\top) \in \mathbb{R}^{d+1}$, and the nonlinearity σ is ReLU. We will keep such an order in θ to view it as a vector. The input $x \in \mathbb{R}^d$ is drawn uniformly from a unit sphere $\mathcal{S}^{d-1}$. The parameters are trained by GD subjected to L_2 population loss, as

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta_t), \quad L(\theta_t) = \mathbb{E}_{x \in \mathcal{S}^{d-1}} \big(f(x;\theta_t) - y\big)^2.$$

We generate labels from a single teacher neuron function, as $y|x = \sigma(\tilde{w}^\top x)$. Hence $\tilde{w}$ is our target neuron to learn. We denote the angle between w and $\tilde{w}$ as $\alpha \geq 0$. Note that α is set as non-negative because the loss function is symmetric w.r.t. the angle.

From the definition, we have

$$\begin{aligned}\nabla_\theta L &= 2\mathbb{E}_{x \in \mathcal{S}^{d-1}} \big(f(x;\theta) - y\big) \nabla_\theta f(x;\theta) = \mathbb{E}_{x \in \mathcal{S}_+^{d-1}(w)} (vw^\top x - \sigma(\tilde{w}^\top x)) \begin{bmatrix} w^\top x \\ vx \end{bmatrix} \\ &= \frac{1}{d} \begin{bmatrix} v\,\|w\|_2^2 - \frac{\|w\|}{\pi}\big(\sin\alpha + (\pi - \alpha)\cos\alpha\big)\,\|\tilde{w}\| \\ v^2 w - \frac{v}{\pi}(\pi - \alpha + \frac{1}{2}\sin 2\alpha) \cdot \tilde{w} - \frac{v}{\pi}(-\frac{1}{2}\cos 2\alpha + \frac{1}{2})\,\|\tilde{w}\|\,\tilde{w}_\perp \end{bmatrix},\end{aligned}$$

where we denote $\tilde{w}_\perp$ as the unit-length orthogonal residual of w after projecting onto $\tilde{w}$ and $\mathcal{S}_+^{d-1}(w)$ is the half-sphere where $\langle w, x\rangle \geq 0$. Consider the Hessian as

$$H \triangleq \begin{bmatrix} \partial_v^2 L & \partial_w \partial_v L \\ \partial_v \partial_w L & \partial_w^2 L \end{bmatrix} = \mathbb{E}_{x \in \mathcal{S}_+^{d-1}(w)} \begin{bmatrix} w^\top x x^\top w & \cdots \\ (vw^\top x - \sigma(\tilde{w}^\top x))x + vxx^\top w & v^2 x x^\top \end{bmatrix} \tag{9}$$

$$\overset{\text{if } vw = \tilde{w}}{=\!=\!=} \frac{1}{d} \begin{bmatrix} \|w\|^2 & vw^\top \\ vw & v^2 \mathbb{I} \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}. \tag{10}$$

Hence, in the global minima manifold where $vw = \tilde{w}$, the eigenvalues of the Hessian are $\lambda_1 = \frac{\|w\|^2 + v^2}{d}, \lambda_{2\ldots d} = \frac{v^2}{d}, \lambda_{d+1} = 0$. Therefore, the largest eigenvalue λ_1 measures the inbalance between the two layers again as $\lambda_1 = \frac{(\|w\| - v)^2 + 2\|\tilde{w}\|}{d}$ similar to the 2-D case in (6). So we would like to investigate where GD converges if $\eta > \frac{2}{2\|\tilde{w}\|/d} = d/\|\tilde{w}\|$. Note that a key difference between the current case and the previous 2-D analysis is that the current one includes a neuron as a vector and a nonlinear ReLU unit.

Before diving into the convergence analysis, let's simplify the setting reasonably. From the second row of $\nabla_\theta L$, which is $\nabla_w L$, it is obvious that updates of w always stay in the plane spanned by $\tilde{w}$ and $w^{(0)}$. Hence, this problem can be simplified to three variables (v, w_x, w_y) with the target neuron $\tilde{w} = [1, 0]$. The three variables stand for

$$v^{(t)} := v^{(t)}, \quad w_x^{(t)} := \mathrm{proj}_{\tilde{w}}\, w^{(t)}, \quad w_y^{(t)} := \mathrm{proj}_{\tilde{w}_\perp}\, w^{(t)} = \sqrt{\big\|w^{(t)}\big\|^2 - (w_x^{(t)})^2}.$$

Denote the learning rate $\eta = Kd$, with $K > 1$. Then the update rule is, following GD, omitting the superscripts of time t in RHS,

$$v^{(t+1)} := v^{(t)} + K w_x \left[(-v w_x + 1) - \frac{1}{\pi}\left(\alpha - \frac{w_x w_y}{\|w\|^2}\right)\right] + K w_y^2 \left(-v + \frac{w_y}{\pi\,\|w\|^2}\right), \tag{11}$$

$$w_x^{(t+1)} := w_x^{(t)} + K v \left[(-v w_x + 1) - \frac{1}{\pi}\left(\alpha - \frac{w_x w_y}{\|w\|^2}\right)\right], \tag{12}$$

$$w_y^{(t+1)} := \left| w_y^{(t)} \cdot \left[1 + K\left(-v^2 + \frac{v w_y}{\pi\,\|w\|^2}\right)\right]\right|. \tag{13}$$

We keep w_y as nonnegative because the loss L is invariant to its sign and our previous notation $\alpha \geq 0$ requires a non-negative w_y. Then we present a convergence result as follows.

Theorem 4. *In the above setting, consider a teacher neuron* $\tilde{w} = [1, 0]$ *and set the learning rate* $K \in (1, 1.1]$. *Initialize the student as* $\left\|w^{(0)}\right\| = v^{(0)} \triangleq \epsilon \in (0, 0.10]$ *and* $\langle w^{(0)}, \tilde{w}\rangle \geq 0$. *Then, for* $t \geq T_1 + 4$, $w_y^{(t)}$ *decays as*

$$w_y^{(t)} < 0.1 \cdot (1 - 0.030K)^{t - T_1 - 4}, \quad T_1 \leq \left\lceil \log_{2.56} \frac{1.35}{\pi\beta^2} \right\rceil, \quad \beta = \left(1 + \frac{1.1}{\pi}\right)\epsilon.$$

Proof sketch The details of proof are presented in the Appendix G. The proof is divided into two stages, depending on whether w_y grows or not. The key is that the sign of Δw_y aligns with the sign of

$$\Delta w_y \propto -v w_x + \frac{1}{\pi} \frac{\frac{w_y}{w_x}}{1 + (\frac{w_y}{w_x})^2}, \tag{14}$$

where the second term of RHS is bounded in $[0, \frac{1}{2\pi}]$. In stage 1 where vw_x is relatively small, we show the growth ratio of w_y is smaller than w_x and vw_x, resulting in an upper bound of number of iterations for vw_x to reach $\frac{1}{2\pi}$, so $\max(w_y)$ is bounded too. Although the initialization is balanced as $v^{(0)} = \left\|w^{(0)}\right\|$ for simplicity of proof, $v - w_x$ is also bounded at the end of stage 1. From the beginning of stage 2, thanks to the relatively narrow range of K, we are able to compute the bounds of three variables (including $v - w_x$, vw_x and w_y) and they turn out to fall into a basin in the parameter space after four iterations. In this basin, w_y decays exponentially with a linear rate of 0.97 at most. □

With the guarantee of w_y decaying in the above theorem, the dynamics of the single-neuron ReLU network is getting closer to the 2-D case in Section 4.2. Actually, in Theorem 4, $w_y < 0.1$ after $t \geq T_1 + 4$ implies, in (11, 12), we have

$$\frac{1}{\pi}\left(\alpha - \frac{w_x w_y}{\|w\|^2}\right) \in [0, 2.6 \times 10^{-4}], \quad K w_y^2 \left(-v + \frac{w_y}{\pi \|w\|^2}\right) \in [-1.45 \times 10^{-2}, 0], \tag{15}$$

while other terms are in the order of $\Theta(1)$.

To summarize, the single-neuron model goes through three phases of training dynamics, with a intialization of the angle as $\frac{\pi}{2}$ at most. First, the angle decreases monotonically but, due to the growth of norms, the absolute deviation w_y still increases. Meanwhile, the inbalance $v - w_x$ stays in a bounded level. Second, w_y starts to decrease and the parameters fall into a basin within four steps. Third, in the basin, w_y decreases exponentially and, after w_y at a reasonable low level, the model approximately follows the dynamic of the 2-D case and the inbalance $v - w_x$ decreases as well. The model converges to a period-2 orbit as in the 1-D case.

6 Numerical Experiments

In this section, we provide numerical experiments to verify our theorems.

1-D functions. As discussed in the Section 4.1, we have $f(x) = \frac{1}{4}(x^2 - 1)^2$ satisfying the condition in Theorem 1 and $g(x) = 2\sin(x)$ satisfying Lemma 1, so we estimate that both f and g allow stable oscillation around the local minima. It turns out GD stably oscillates around the local minima on both functions, when $\eta > \frac{2}{f''(\bar{x})}$ slightly, as shown in Figure 1.

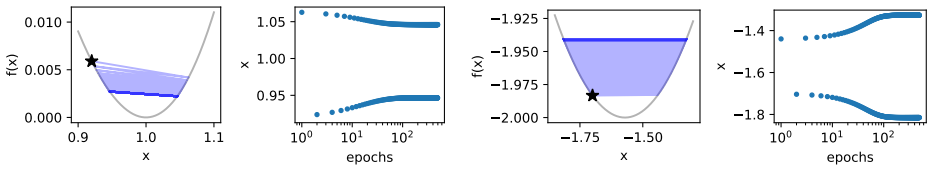
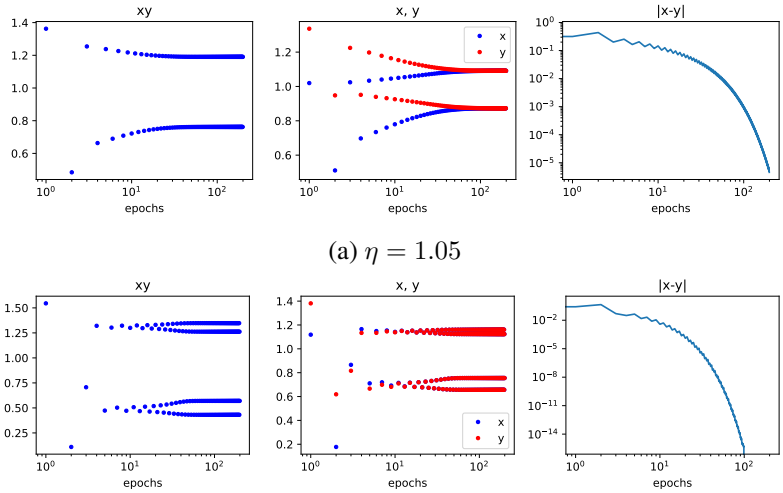


Figure 1: Running GD around the local minima of $f(x) = \frac{1}{4}(x^2 - 1)^2$ (left) and $f(x) = 2\sin(x)$ (right) with learning rate $\eta = 1.01 > \frac{2}{f''(\bar{x})} = 1$. Stars denote the start points. It turns out both functions allow stable oscillation around the local minima.

2-D function. As discussed in the Section 4.2, on the function $f(x, y) = \frac{1}{2}(xy - 1)^2$, we estimate that $|x - y|$ decays to 0 when $\eta \in (1, 1.5)$, as shown in Figure 2. Since it achieves a perfect balance, the two parameters follows convergence of the corresponding 1-D function $f(x) = \frac{1}{4}(x^2 - 1)^2$. As shown in Figure 2, xy with $\eta = 1.05$ converges to a period-2 orbit, as stated in the 1-D discussion of Theorem 2 while xy with $\eta = 1.25$ converges to a period-4 orbit, which is out of our range in the theorem. But still it falls into the range for balance in Theorem 3.



(a) $\eta = 1.05$

(b) $\eta = 1.25$

Figure 2: Running GD on $f(x, y) = \frac{1}{2}(xy - 1)^2$ with learning rate $\eta = 1.05$ (top) and $\eta = 1.25$ (bottom). When $\eta = 1.05$, it converges to a period-2 orbit. When $\eta = 1.25$, it converges to a period-4 orbit. In both cases, $|x - y|$ decays sharply.

Two-layer single-neuron model. As discussed in the Section 5, with a learning rate $\eta \in (d, 1.1d]$, a single-neuron network $f(x) = v \cdot \sigma(w^\top x)$ is able to align with the direction of the teacher neuron under population loss. We train such a model in empirical loss on 1000 data points uniformly sampled from a sphere $\mathcal{S}^1$, as shown in Figure 3. The student neuron is initialized orthogonal to the teacher neuron. In the end of training, w_y decays to a small value before the inbalance $|v - w_x|$ decays sharply, which verifies our argument in Section 5. With a small w_y, this nonlinear problem degenerates to a 2-D problem on v, w_x. Then, the balanced property makes it align with the 1-D problem where v and w_x converge to a period-2 orbit. Note that the small residuals of $|v - w_x|$ and w_y are due to the difference between population loss and empirical loss.

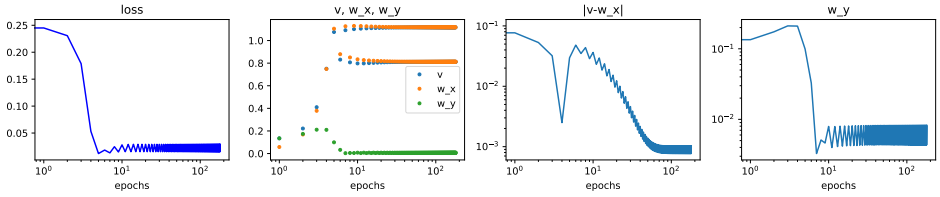


Figure 3: Running GD in the teacher-student setting with learning rate $\eta = 2.2 = 1.1d$, trained on 1000 points uniformly sampled from sphere $\mathcal{S}^1$ of $\|x\| = 1$. The teacher neuron is $\tilde{w} = [1, 0]$ and the student neuron is initialized as $w^{(0)} = [0, 0.1]$ with $v^{(0)} = 0.1$.

7 Conclusions

In this work, we investigate gradient descent with a large step size that crosses the threshold of local stability. In the low dimensional setting, we provide conditions on high-order derivatives that allow stable oscillation around local minima. Meanwhile, we establish a balancing effect whereby

inbalanced initialization ends in balance, where the loss landscape is flatter. For a two-layer single-neuron ReLU network, we prove its convergence to align with the teacher neuron under population loss.

While these are encouraging results that contribute to the growing understanding of gradient descent beyond the Edge of Stability, our analysis suffers from important limitations that require further work. First, our analysis is inherently low-dimensional, and we are able to apply it to the teacher-student ReLU setting thanks to the rotational symmetry of the population loss. An important item for future work is therefore to extend it to general dimensions, which will enable the analysis of empirical landscapes as well as multiple neurons. Next, the understanding of the implicit bias of GD in the large-learning rate regime won't be complete without integrating the noise, either in the classic SGD sense or in the labels, as done in Damian et al. (2021); Li et al. (2021).

REFERENCES

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. *arXiv preprint arXiv:2204.01050*, 2022.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on edge of stability in deep learning. *arXiv preprint arXiv:2205.09745*, 2022.

David Barrett and Benoit Dherin. Implicit gradient regularization. In *International Conference on Learning Representations*, 2020.

Nils Bjorck, Carla P Gomes, Bart Selman, and Kilian Q Weinberger. Understanding batch normalization. *Advances in neural information processing systems*, 31, 2018.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2020.

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

Lijun Ding, Dmitriy Drusvyatskiy, and Maryam Fazel. Flat minima generalize for low-rank matrix recovery. *arXiv preprint arXiv:2203.03756*, 2022.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in Neural Information Processing Systems*, 31, 2018.

Omer Elkabetz and Nadav Cohen. Continuous vs. discrete optimization of deep neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Justin Gilmer, Behrooz Ghorbani, Ankush Garg, Sneha Kudugunta, Behnam Neyshabur, David Cardoze, George Dahl, Zachary Nado, and Orhan Firat. A loss curvature perspective on training instability in deep learning. *arXiv preprint arXiv:2110.04369*, 2021.

Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. *Advances in neural information processing systems*, 32, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. In *International Conference on Learning Representations*, 2019.

Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J Geras. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *International Conference on Machine Learning*, pp. 4772–4784. PMLR, 2021.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29, 2021.

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.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pp. 9–48. Springer, 2012.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after sgd reaches zero loss?–a mathematical framework. *arXiv preprint arXiv:2110.06914*, 2021.

Chao Ma and Lexing Ying. The sobolev regularization effect of stochastic gradient descent. *arXiv preprint arXiv:2105.13462*, 2021.

Chao Ma, Lei Wu, and Lexing Ying. The multiscale structure of neural network loss functions: The effect on optimization and origin. *arXiv preprint arXiv:2204.11326*, 2022.

Yu Nesterov. Introductory lectures on convex programming, 1998.

Samuel L Smith, Benoit Dherin, David GT Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. *arXiv preprint arXiv:2101.12176*, 2021.

Gal Vardi and Ohad Shamir. Implicit regularization in relu networks with the square loss. In *Conference on Learning Theory*, pp. 4224–4258. PMLR, 2021.

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

Yuqing Wang, Minshuo Chen, Tuo Zhao, and Molei Tao. Large learning rate tames homogeneity: Convergence and balancing effect. *arXiv preprint arXiv:2110.03677*, 2021.

Tian Ye and Simon S Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. *Advances in Neural Information Processing Systems*, 34, 2021.

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

Appendices

A Proof of Theorem 1 **13**

B Proof of Lemma 1 **13**

C Proof of Lemma 2 **15**

D Proof of Theorem 2 **16**

E Proof of Theorem 3 **20**

F Proof of Lemma 3 **23**

G Proof of Theorem 4 **24**

H Useful lemmas **39**

I Additional Experiments **40**

I.1 2-layer high-dim homogeneous ReLU NNs with planted teacher neurons 40

I.2 3, 4, 5-layer non-homogeneous MLPs on MNIST 42

J Illustration of period-2 and period-4 orbits **45**

A PROOF OF THEOREM 1

Theorem 5 (Restatement of Theorem 1)**.** *Consider a 1-D differentiable function $f(x)$ around a local minima $\bar{x}$. Assume 1) $f^{(3)}(\bar{x}) \neq 0$, 2) $\frac{f^{(3)}(\bar{x})}{f''(\bar{x})} = \mathcal{O}(1)$, 3) $3[f^{(3)}]^2 - f''f^{(4)} > 0$ at $x = \bar{x}$, and 4) all higher order derivatives are bounded. For a start point near $\bar{x}$ as $x_0 = x - \epsilon$ such that $\epsilon \cdot f^{(3)}(\bar{x}) > 0$ and ϵ is sufficiently small, then there exists a learning rate η with which GD starting from x_0 will back to it in two steps, and*

$$\frac{2}{f''(\bar{x})} < \eta < \frac{2}{f''(\bar{x}) - \epsilon \cdot f^{(3)}(\bar{x})}.$$

Proof. For simplicity, we assume $f^{(3)}(\bar{x}) > 0$. Imagine a starting point $x_0 = \bar{x} - \epsilon, \epsilon > 0$. We omit $f'(\bar{x}), f''(\bar{x}), f^{(3)}(\bar{x}), f^{(4)}(\bar{x})$ as $f', f'', f^{(3)}, f^{(4)}$. After running two steps of gradient descent, we have

$$\begin{aligned}
x_0 &= \bar{x} - \epsilon, \\
f'(x_0) &= f' - f''\epsilon + \frac{1}{2}f^{(3)}\epsilon^2 - \frac{1}{6}f^{(4)}\epsilon^3 + \mathcal{O}(\epsilon^4) \\
&= -f''\epsilon + \frac{1}{2}f^{(3)}\epsilon^2 - \frac{1}{6}f^{(4)}\epsilon^3 + \mathcal{O}(\epsilon^4), \\
x_1 &= x_0 - \eta f'(x_0) = \bar{x} - \epsilon - \eta\big(-f''\epsilon + \frac{1}{2}f^{(3)}\epsilon^2 - \frac{1}{6}f^{(4)}\epsilon^3\big) + \mathcal{O}(\epsilon^4), \\
f'(x_1) &= f'' \cdot (x_1 - \bar{x}) + \frac{1}{2}f^{(3)} \cdot (x_1 - \bar{x})^2 + \frac{1}{6}f^{(4)} \cdot (x_1 - \bar{x})^3 + \mathcal{O}(\epsilon^4), \\
x_2 &= x_1 - \eta f'(x_1), \\
\frac{x_2 - x_0}{\eta} &= -\left(-f''\epsilon + \frac{1}{2}f^{(3)}\epsilon^2 - \frac{1}{6}f^{(4)}\epsilon^3\right) - f'' \cdot \left(-\epsilon - \eta\big(-f''\epsilon + \frac{1}{2}f^{(3)}\epsilon^2 - \frac{1}{6}f^{(4)}\epsilon^3\big)\right) \\
&\quad - \frac{1}{2}f^{(3)}\left(-\epsilon - \eta\big(-f''\epsilon + \frac{1}{2}f^{(3)}\epsilon^2 - \frac{1}{6}f^{(4)}\epsilon^3\big)\right)^2 - \frac{1}{6}f^{(4)} \cdot (-\epsilon - \eta(-f''\epsilon))^3 + \mathcal{O}(\epsilon^4) \\
&= (2f'' - \eta f''f'')\,\epsilon + \left(-\frac{1}{2}f^{(3)} + \frac{1}{2}\eta f''f^{(3)} - \frac{1}{2}f^{(3)}(-1 + \eta f'')^2\right)\epsilon^2 \\
&\quad + \left(\frac{1}{6}f^{(4)} - \frac{1}{6}\eta f''f^{(4)} + \frac{1}{2}(-1 + \eta f'')\eta f^{(3)}f^{(3)} - \frac{1}{6}(-1 + \eta f'')^3 f^{(4)}\right)\epsilon^3 + \mathcal{O}(\epsilon^4).
\end{aligned}$$

When $\eta = \frac{2}{f''}$, it holds

$$\frac{x_2 - x_0}{\eta} = \left(\frac{1}{2}\eta f^{(3)}f^{(3)} - \frac{1}{3}f^{(4)}\right)\epsilon^3 + \mathcal{O}(\epsilon^4), \tag{16}$$

which would be positive if $f^{(4)}$ is bounded.

When $\eta = \frac{2}{f'' - \epsilon \cdot f^{(3)}}$ then $\eta f'' = 2 + 2\frac{f^{(3)}}{f''}\epsilon + \mathcal{O}(\epsilon^2)$, it holds

$$\frac{x_2 - x_0}{\eta} = -2f^{(3)}\epsilon^2 + \left(-\frac{1}{2}f^{(3)} + f^{(3)} - \frac{1}{2}f^{(3)}\right)\epsilon^2 + \mathcal{O}(\epsilon^3) = -2f^{(3)}\epsilon^2 + \mathcal{O}(\epsilon^3), \tag{17}$$

which is negative when ϵ is small.

Therefore, there exists a learning rate $\eta \in (\frac{2}{f''}, \frac{2}{f'' - \epsilon \cdot f^{(3)}})$ such that $x_2 = x_0$. □

B PROOF OF LEMMA 1

Lemma 4 (Restatement of Lemma 1)**.** *Consider a 1-D differentiable function $f(x)$ around a local minima $\bar{x}$. Assume 1) the lowest order non-zero derivative of f (except the f'') at $\bar{x}$ is $f^{(k)}(\bar{x})$, 2) $\frac{f^{(k)}(\bar{x})}{f''(\bar{x})} = \mathcal{O}(1)$, 3) all higher order derivatives are bounded, and 4) $k \geq 4$. Then, from a starting point near $\bar{x}$ as $x_0 = \bar{x} - \epsilon$ with $\epsilon > 0$ sufficiently small, and*

1. *if k is odd and $f^{(k+1)}(\bar{x}) < 0$ (and reasonably assuming $f^{(k)} > 0$ as remarked below), then there exists $\eta \in (\frac{2}{f''}, \frac{2}{f''-f^{(k)}\epsilon^{k-2}})$,*

2. *if k is even and $f^{(k)}(\bar{x}) < 0$, then there exists $\eta \in (\frac{2}{f''}, \frac{2}{f''+f^{(k)}\epsilon^{k-2}})$,*

such that: GD can stably oscillate in the neighborhood of $\bar{x}$ with $\eta > 2/f''(\bar{x})$ by a sufficiently small margin.

Remark. The following proof for the odd case is for both ϵ and $f^{(k)}(\bar{x})$ are positive. However, it only needs a little change if $f^{(k)} < 0$. If k is odd and $f^{(k)} < 0$, there are two alternative ways to complete the proof:

1. Mirror the function as $\tilde{f}(x) \coloneqq f(-x)$. Then in the odd case of $\tilde{f}(x)$, it has $\tilde{f}^k(-\bar{x}) > 0$. Then the same discussion starting from $\bar{x}_0 = -\bar{x} - \epsilon$ with $\epsilon > 0$ is enough.

2. Set $\epsilon < 0$ and very close to 0. Then the signs of (18, 19) would flip to negative and positive, so the result still holds.

Proof. (1) If k is odd, assuming $f^{(k)} > 0$ for simplicity, we have

$$
\begin{aligned}
x_0 &= \bar{x} - \epsilon, \\
f'(x_0) &= -f''\epsilon + \frac{1}{(k-1)!}f^{(k)}\epsilon^{k-1} - \frac{1}{k!}f^{(k+1)}\epsilon^k + \mathcal{O}(\epsilon^{k+1}), \\
x_1 &= x_0 - \eta f'(x_0) = \bar{x} - \epsilon + \eta f''\epsilon - \frac{1}{(k-1)!}\eta f^{(k)}\epsilon^{k-1} + \frac{1}{k!}\eta f^{(k+1)}\epsilon^k + \mathcal{O}(\epsilon^{k+1}), \\
f'(x_1) &= f'' \cdot (x_1 - \bar{x}) + \frac{1}{(k-1)!}f^{(k)} \cdot (x_1 - \bar{x})^{k-1} + \frac{1}{k!}f^{(k+1)} \cdot (x_1 - \bar{x})^k + \mathcal{O}(\epsilon^{k+1}), \\
\frac{x_2 - x_0}{\eta} &= \frac{x_1 - \eta f'(x_1) - x_0}{\eta} = -f'(x_0) - f'(x_1) \\
&= (2f'' - \eta f''f'')\,\epsilon \\
&\quad + \left(-\frac{1}{(k-1)!}f^{(k)} + \frac{1}{(k-1)!}\eta f''f^{(k)} - \frac{1}{(k-1)!}f^{(k)} \cdot (-1 + \eta f'')^{k-1}\right)\epsilon^{k-1} \\
&\quad + \left(\frac{1}{k!}f^{k+1} - \frac{1}{k!}\eta f''f^{(k+1)} - \frac{1}{k!}f^{(k+1)} \cdot (-1 + \eta f'')^k\right)\epsilon^k + \mathcal{O}(\epsilon^{k+1})
\end{aligned}
$$

When $\eta = \frac{2}{f''}$, it holds

$$
\frac{x_2 - x_0}{\eta} = -\frac{2}{k!}f^{(k+1)}\epsilon^k + \mathcal{O}(\epsilon^{k+1}), \tag{18}
$$

which is positive if $f^{(k+1)} < 0$ with $k+1$ as even.

When $\eta = \frac{2}{f''-f^{(k)}\epsilon^{k-2}}$ then $\eta f'' = 2 + 2\frac{f^{(k)}}{f''}\epsilon^{k-2} + \mathcal{O}(\epsilon^{2k-4})$, then it holds

$$
\frac{x_2 - x_0}{\eta} = -2f^{(k)}\epsilon^{k-1} + \mathcal{O}(\epsilon^k), \tag{19}
$$

which is negative.

Therefore, there exists a learning rate $\eta \in (\frac{2}{f''}, \frac{2}{f''-f^{(k)}\epsilon^{k-2}})$ such that $x_2 = x_0$.

(2) If k is even, we have

$$
\begin{aligned}
x_0 &= \bar{x} - \epsilon, \\
f'(x_0) &= -f''\epsilon - \frac{1}{(k-1)!} f^{(k)} \epsilon^{k-1} + \mathcal{O}(\epsilon^k), \\
x_1 &= x_0 - \eta f'(x_0) = \bar{x} - \epsilon + \eta f''\epsilon + \frac{1}{(k-1)!} \eta f^{(k)} \epsilon^{k-1} + \mathcal{O}(\epsilon^k), \\
f'(x_1) &= f'' \cdot (x_1 - \bar{x}) + \frac{1}{(k-1)!} f^{(k)} \cdot (x_1 - \bar{x})^{k-1} + \mathcal{O}(\epsilon^k), \\
\frac{x_2 - x_0}{\eta} &= \frac{x_1 - \eta f'(x_1) - x_0}{\eta} = -f'(x_0) - f'(x_1) \\
&= (2f'' - \eta f'' f'')\, \epsilon \\
&\quad + \left(\frac{1}{(k-1)!} f^{(k)} - \frac{1}{(k-1)!} \eta f'' f^{(k)} - \frac{1}{(k-1)!} (-1 + \eta f'')^{k-1} \right) \epsilon^{k-1} + \mathcal{O}(\epsilon^k).
\end{aligned}
$$

When $\eta = \frac{2}{f''}$, it holds

$$
\frac{x_2 - x_0}{\eta} = -\frac{2}{(k-1)!} f^{(k)} \epsilon^{k-1} + \mathcal{O}(\epsilon^k),
$$

which is positive.

When $\eta = \frac{2}{f'' + c \cdot f^{(k)} \epsilon^{k-2}}$ with $c > 0$ as some constant implying $\eta f'' = 2(1 - c \frac{f^{(k)}}{f''} \epsilon^{k-2}) + \mathcal{O}(\epsilon^{2k-4})$, then it holds

$$
\frac{x_2 - x_0}{\eta} = 2 \left(c - \frac{1}{(k-1)!} \right) f^{(k)} \epsilon^{k-1} + \mathcal{O}(\epsilon^k),
$$

which would be negative, if we set the constant as $c > \frac{1}{(k-1)!}$ by a sufficient margin, such as $c = 1$.

Therefore, there exists a learning rate $\eta \in (\frac{2}{f''}, \frac{2}{f'' + f^{(k)} \epsilon^{k-2}})$ such that $x_2 = x_0$. □

Corollary 1. *$f(x) = \sin(x)$ allows stable oscillation around its local minima $\bar{x}$.*

Proof. Its lowest order nonzero derivative (expect f'') is $f^{(4)} \bar{x} = \sin(\bar{x}) = -1 < 0$ and the order 4 is even. Then Lemma 1 gives the result. □

C Proof of Lemma 2

Lemma 5 (Restatement of Lemma 2)**.** *Consider a 1-D function $g(x)$ with the a target value y. Then define the loss function f as $f(x) = (g(x) - y)^2$. Assuming 1) g' is not zero when $g(\bar{x}) = y$, 2) all higher order derivatives are well bounded, then it satisfies the condition in Theorem 1 or Lemma 1 to allow stable oscillation around $\bar{x}$. More precisely, it holds if*

$$
g'(\bar{x}) g^{(3)}(\bar{x}) < 6[g''(\bar{x})]^2.
$$

Proof. From the definition, we have

$$
f''(x) = 2[g(x) - y] g''(x) + 2[g'(x)]^2, \tag{20}
$$

$$
f^{(3)}(x) = 2[g(x) - y] g^{(3)}(x) + 6 g''(x) g'(x), \tag{21}
$$

$$
f^{(4)}(x) = 2[g(x) - y] g^{(4)}(x) + 6 g''(x) g''(x) + 8 g'(x) g^{(3)}(x). \tag{22}
$$

Then at the global minima where $g(x) = y$, we have $f''(x) = 2[g'(x)]^2$ and $f^{(3)}(x) = 6 g''(x) g'(x)$. If we assume y is not a trivial value for $g(x)$, which means $g'(x) \neq 0$ at the minima, and g is not

linear around the minima (implies $g'' \neq 0$), then f satisfies the condition in Theorem 1 to allow local stable oscillation. More concretely, we need $\frac{1}{f''} f^{(3)} f^{(3)} - \frac{1}{3} f^{(4)} > 0$ as in (16), hence it requires

$$\frac{1}{2g'(x)g'(x)} 36 g''(x)g''(x)g'(x)g'(x) - \frac{1}{3}\left(6g''(x)g''(x) + 8g'(x)g^{(3)}(x)\right) > 0 \tag{23}$$

$$6g''(x)g''(x) > g'(x)g^{(3)}(x), \tag{24}$$

where $g^{(3)}$ needs to be well bounded.

Meanwhile, if $g'(x) \neq 0$ and $g'' = 0$ at the minima, it satisfies the condition for Lemma 1 with $k = 4$, because $f^{(3)} = 0$ and $f^{(4)} < 0$ due to (22, 24) □

Corollary 2. *$f(x) = (x^2 - 1)^2$ allows stable oscillation around the local minima $\bar{x} = 1$.*

Proof. With $g(x) = x^2$, it has $g'(1) = 2 \neq 0, g''(1) = 2 \neq 0$. All higher order derivatives of g are zero. Then Lemma 2 gives the result. □

Corollary 3. *$f(x) = (\sin(x) - y)^2$ allows stable oscillation around the local minima $\bar{x} = \arcsin(y)$ with $y \in (-1, 1)$.*

Proof. With $g(x) = \sin(x)$, it has $g'(\bar{x}) = \cos(\bar{x}) \neq 0, g^{(3)}(\bar{x}) = -\cos(\bar{x})$. We have $g^{(3)}(\bar{x})$ is bounded as $g'g^{(3)} - 6[g'']^2 = -\cos^2(\bar{x}) - 6\sin^2(\bar{x}) < 0$. Then Lemma 2 gives the result. □

Corollary 4. *$f(x) = (\tanh(x) - y)^2$ allows stable oscillation around the local minima $\bar{x} = \tanh^{-1}(y)$ with $y \in (-1, 1)$.*

Proof. With $g(x) = \tanh(x)$, it has $g'(\bar{x}) = \text{sech}^2(\bar{x}) \neq 0$, and $g^{(3)}(\bar{x}) = -2\text{sech}^4(\bar{x}) + 4\text{sech}^2(\bar{x})\tanh^2(\bar{x})$ is bounded as

$$g'g^{(3)} - 6[g'']^2 = -2\text{sech}^6 + 4\text{sech}^4\tanh^2 - 24\text{sech}^4\tanh^2 = -2\text{sech}^6 - 20\text{sech}^4\tanh^2 < 0.$$

Then Lemma 2 gives the result. □

D Proof of Theorem 2

Theorem 6 (Restatement of Theorem 2)**.** *For $f(x) = \frac{1}{4}(x^2 - \mu)^2$, consider GD with $\eta = K \cdot \frac{1}{\mu}$ where $1 < K \leq 1.121$. The process starts from any point $0 < x < \sqrt{\mu}$. Then it converges to an orbit of period 2, except for a measure-zero initialization where it converges to $\sqrt{\mu}$.*

Proof. Assume the 2-period orbit is $(\bar{x}_0, \bar{x}_1)$, which means

$$\bar{x}_1 = \bar{x}_0 - \eta \cdot f'(\bar{x}_0) = \bar{x}_0 + \eta \cdot (\mu - \bar{x}_0^2)\bar{x}_0,$$
$$\bar{x}_0 = \bar{x}_1 - \eta \cdot f'(\bar{x}_1) = \bar{x}_1 + \eta \cdot (\mu - \bar{x}_1^2)\bar{x}_1.$$

First, we show the existence and uniqueness of such an orbit when $K \in (1, 1.5]$ via solving a high-order equation, some roots of which can be eliminated. Second, we present a result of local convergence near $(\bar{x}_0, \bar{x}_1)$, where the residual error to the orbit decays exponentially when $K \in (1, 1.121)$. Third, we generalize the analysis into global convergence by defining a special interval I. GD starting from any point following our assumption will enter I in some steps, and any point in I will back to this interval after two steps of iteration. Finally, any point in I will converge to the orbit $(\bar{x}_0, \bar{x}_1)$.

Before diving into the proof, we briefly show it always holds $x > 0$ under our assumption. If $x_{t-1} > 0$ and $x_t \leq 0$, the GD rule reveals $\eta(\mu - x_{t-1}^2) \leq -1$ which implies $x_{t-1}^2 \geq \mu + \frac{1}{\eta}$. However, the maximum of $x + \eta(\mu - x^2)x$ on $x \in (0, \sqrt{\mu + \frac{1}{\eta}})$ is achieved when $x^2 = \frac{1}{3}(\mu + \frac{1}{\eta})$ so the maximum value is $\sqrt{\frac{1}{3}(\mu + \frac{1}{\eta})}(\frac{2}{3} + \frac{2}{3}\eta\mu) \leq 1.4\sqrt{\frac{1}{3}(\mu + \frac{1}{\eta})} < \sqrt{\mu + \frac{1}{\eta}}$. As a result, it always holds $x > 0$.

Part I. Existence and uniqueness of $(\bar{x}_0, \bar{x}_1)$.

In this part, we simply denote both $\bar{x}_0, \bar{x}_1$ as x_0. This means x_0 in all formulas in this part can be interpreted as $\bar{x}_0$ and $\bar{x}_1$. Then the GD update rule tells, for the orbit in two steps,

$$x_0 \mapsto x_1 \coloneqq x_0 + \eta(\mu - x_0^2)x_0,$$
$$x_1 \mapsto x_0 = x_1 + \eta(\mu - x_1^2)x_1,$$

which means

$$0 = \eta(\mu - x_0^2)x_0 + \eta\left(\mu - \left(x_0 + \eta(\mu - x_0^2)x_0\right)^2\right)\left(x_0 + \eta(\mu - x_0^2)x_0\right),$$
$$0 = \mu - x_0^2 + \left(\mu - \left(x_0 + \eta(\mu - x_0^2)x_0\right)^2\right)\left(1 + \eta(\mu - x_0^2)\right).$$

Denote $z \coloneqq 1 + \eta(\mu - x_0^2)$, it is equivalent to

$$\begin{aligned} 0 &= \mu - x_0^2 + (\mu - z^2 x_0^2)z = (z+1)(-x_0^2 z^2 + x_0^2 z + \mu - x_0^2) \\ &= (z+1)\left(-x_0^2(z - \frac{1}{2})^2 + \mu - \frac{3}{4}x_0^2\right). \end{aligned}$$

If $z + 1 = 0$, it means $x_1 = -x_0$ which is however out of the range of our discussion on the $x > 0$ domain. So we require $-x_0^2(z - \frac{1}{2})^2 + \mu - \frac{3}{4}x_0^2 = 0$. To ensure the existence of solutions z, it is natural to require

$$\mu - \frac{3}{4}x_0^2 \geq 0.$$

Then, the solutions are

$$z = \frac{1}{2} \pm \sqrt{\frac{\mu}{x_0^2} - \frac{3}{4}}.$$

However, $z = \frac{1}{2} - \sqrt{\frac{\mu}{x_0^2} - \frac{3}{4}}$ can be ruled out. If it holds, $\eta(\mu - x_0^2) = z - 1 < -\frac{1}{2}$ which means $x_0^2 > \mu + \frac{1}{2\eta}$. Since we restrict $\eta\mu \in (1, 1.121]$, it tells $x_0^2 > \mu(1 + \frac{1}{1.242})$ contradicting with $\mu \geq \frac{3}{4}x_0^2$.

Hence, $z = \frac{1}{2} + \sqrt{\frac{\mu}{x_0^2} - \frac{3}{4}}$ is the only reasonable solution, which is saying

$$\eta(\mu - x_0^2) = -\frac{1}{2} + \sqrt{\frac{\mu}{x_0^2} - \frac{3}{4}}.$$

Given a certain η, the above expression is a third-order equation of x_0^2 to solve. Apparently $x_0^2 = \mu$ is one trivial solution, since for any learning rate, the gradient descent stays at the global minimum. Then the two other solutions are exactly the orbit $(\bar{x}_0, \bar{x}_1)$, if the equation does have three different roots. This also guarantees the uniqueness of such an orbit.

Assuming $x_0^2 \neq \mu$, the above expression can be reformulated as

$$\eta = \frac{1}{x_0^2\left(\sqrt{\frac{\mu}{x_0^2} - \frac{3}{4}} + \frac{1}{2}\right)}. \tag{25}$$

One necessary condition for existence is $\mu \geq \frac{3}{4}x_0^2$. Note that here x_0 can be both $\bar{x}_0, \bar{x}_1$, one of which is larger than $\sqrt{\mu}$. For simplicity, we assume $\bar{x}_0 < \sqrt{\mu} < \bar{x}_1$. Since η from Eq(25) is increasing with x_0^2 when $\mu < x_0^2$, let $x_0^2 = \frac{4}{3}\mu$ and achieve the upper bound as

$$\eta\mu \leq \frac{3}{2}, \tag{26}$$

which is satisfied by our assumption $1 < \eta\mu \leq 1.121$.

Therefore, we have shown the existence and uniqueness of a period-2 orbit.

Part II. Local convergence near $(\bar{x}_0, \bar{x}_1)$.

Consider a point near $\bar{x}_0$ as $x_0 = \bar{x}_0 + \epsilon$, we have

$$\begin{aligned} x_1 &\coloneqq x_0 + \eta\left(\mu - x_0^2\right) x_0 = \bar{x}_0 + \epsilon + \eta\left(\mu - \left(\bar{x}_0 + \epsilon\right)^2\right)\left(\bar{x}_0 + \epsilon\right) \\ &= \bar{x}_1 + \epsilon + \eta\left(\mu - \bar{x}_0^2\right)\epsilon + \eta\left(-2\bar{x}_0\epsilon^2 + \epsilon^3\right) + \eta\left(-2\bar{x}_0^2\epsilon - \bar{x}_0\epsilon^2\right) \\ &= \bar{x}_1 + \epsilon\left(1 + \eta\left(\mu - 3\bar{x}_0^2\right)\right) + \mathcal{O}(\epsilon^2), \end{aligned}$$

where we use the fact that $\bar{x}_1 = \bar{x}_0 + \eta\left(\mu - \bar{x}_0^2\right)\bar{x}_0$.

Denote $\epsilon_1 \coloneqq x_1 - \bar{x}_1$, then the above equation is equivalent to

$$\epsilon_1 = \epsilon\left(1 + \eta\left(\mu - 3\bar{x}_0^2\right)\right) + \mathcal{O}(\epsilon^2).$$

Similarly, with $\epsilon_2 \coloneqq x_2 - \bar{x}_0$, we have

$$\epsilon_2 = \epsilon_1\left(1 + \eta\left(\mu - 3\bar{x}_1^2\right)\right) + \mathcal{O}(\epsilon_1^2).$$

To have $|\epsilon_2| < |\epsilon|$, ignoring the high order terms of errors, we need

$$\left|\left(1 + \eta\left(\mu - 3\bar{x}_0^2\right)\right)\left(1 + \eta\left(\mu - 3\bar{x}_1^2\right)\right)\right| < 1. \tag{27}$$

Reformulating Eq(25), we have

$$\begin{aligned} \mu x^2 - \frac{3}{4}x^4 &= \left(\frac{1}{\eta} - \frac{1}{2}x^2\right)^2 \\ x^4 - \left(\mu + \frac{1}{\eta}\right)x^2 + \frac{1}{\eta^2} &= 0, \end{aligned}$$

and, by applying Vieta's formulas, we have

$$\begin{aligned} \bar{x}_0^2 + \bar{x}_1^2 &= \mu + \frac{1}{\eta}, \\ \bar{x}_0^2\bar{x}_1^2 &= \frac{1}{\eta^2}. \end{aligned}$$

Combining with Exp(27), it requires

$$\begin{aligned} \left|\left(1 + \eta\left(\mu - 3\bar{x}_0^2\right)\right)\left(1 + \eta\left(\mu - 3\bar{x}_1^2\right)\right)\right| &= \left|(1+\eta\mu)^2 - 3\eta(1+\eta\mu)(\mu + \frac{1}{\eta}) + 9\right| \\ &= \left|-2(1+\eta\mu)^2 + 9\right| \\ &< 1, \end{aligned}$$

which means

$$1 < \eta\mu < \sqrt{5} - 1. \tag{28}$$

Intuitively, for $\sqrt{5} - 1 < \eta\mu < \frac{3}{2}$, the 2-period orbit exists but is probably impossible to reach. Because initialization close to the orbit will diverge from it.

Actually, if we more strictly require

$$0 \leq \left(1 + \eta\left(\mu - 3\bar{x}_0^2\right)\right)\left(1 + \eta\left(\mu - 3\bar{x}_1^2\right)\right) < 1,$$

it is equivalent to require

$$1 < \eta\mu \leq \sqrt{4.5} - 1 \approx 1.121, \tag{29}$$

which is still satisfied by our assumption on $\eta\mu$.

Therefore, we have shown the local convergence near the orbit, with the error term decaying exponentially.

Part III. Global convergence to $(\bar{x}_0, \bar{x}_1)$.

The proof structure is as follows:

1. There exists a special interval $I \coloneqq [x_s, \sqrt{\mu})$ such that any point in I will back to this interval surely after two steps of gradient descent. And $\bar{x}_0 \in I$.
2. Initialized from any point in I, the gradient descent process will surely converge to $\bar{x}_0$ (every two steps of GD).
3. Initialized from any point between 0 and $\sqrt{\mu}$, the gradient descent process will fall into I in some steps.

(III.1) Consider a function $F_\eta(x) = x + \eta(\mu - x^2)x$ performing one step of gradient descent. Since $F'_\eta(x) = 1 + \eta\mu - 3\eta x^2$, we have $F'_\eta(x) > 0$ for $0 < x^2 < \frac{1}{3}\left(\mu + \frac{1}{\eta}\right)$ and $F'_\eta(x) < 0$ otherwise. It is obvious that the threshold has $x_s^2 \coloneqq \frac{1}{3}\left(\mu + \frac{1}{\eta}\right) < \mu$. In the other words, for any point on the right of x_s, GD returns a point in a decreasing manner.

To prove anything further, we would like to restrict $\bar{x}_0 \geq x_s$, which is

$$\bar{x}_0^2 \geq \frac{1}{3}\left(\mu + \frac{1}{\eta}\right) = \frac{1}{3}\left(\mu + \bar{x}_0^2\left(\sqrt{\frac{\mu}{\bar{x}_0^2} - \frac{3}{4}} + \frac{1}{2}\right)\right).$$

Solving this inequality tells

$$\bar{x}_0^2 \geq \frac{3 + \sqrt{2}}{7}\mu. \tag{30}$$

Consequently, by applying Eq(25), we have

$$\eta\mu \leq \sqrt{4.5} - 1 \approx 1.121, \tag{31}$$

coinciding with Exp(29).

With the above discussion of x_s, we are able to define the special internal $I \coloneqq [x_s, \sqrt{\mu})$. Consider a function representing two steps of gradient descent $F_\eta^2(x) \coloneqq F_\eta(F_\eta(x))$. From previous discussion, we know $F_\eta^2(\bar{x}_0) = \bar{x}_0$. Meanwhile, in a small left neighborhood of $\bar{x}_0$, for $x = \bar{x}_0 - \epsilon$, we have

$$-\epsilon < F_\eta^2(\bar{x}_0 - \epsilon) - \bar{x}_0 \leq 0,$$

where both inequalities are from the local convergence analysis of Exp(29). The left inequality is equivalent to $F_\eta^2(\bar{x}_0 - \epsilon) > \bar{x}_0 - \epsilon$. From the uniqueness of $\bar{x}_0$, we know there is no point in $x \in [x_s, \bar{x}_0)$ such that $F_\eta^2(\bar{x}_0) = x$. Considering $F_\eta^2(\cdot)$ is a continuous function, we have

$$F_\eta^2(x) > x, \ \ \forall x \in [x_s, \bar{x}_0).$$

Similarly, we have

$$F_\eta^2(x) < x, \ \ \forall x \in (\bar{x}_0, \sqrt{\mu}).$$

Since $F_\eta(\cdot)$ is decreasing on $[x_s, \infty)$ and $F_\eta(x) > x_s$ for $x \in [x_s, \sqrt{\mu}]$, it is fair to say $F_\eta^2(x)$ is increasing on $x \in [x_s, \sqrt{\mu}]$. A consequent result is that $\bar{x}_0$ is the upper and lower bound for the above two expressions, taking $F^2(\bar{x}_0) = \bar{x}_0$. Formally, we have

$$F_\eta^2(x) \in (x, \bar{x}_0], \ \ \forall x \in [x_s, \bar{x}_0), \tag{32}$$

$$F_\eta^2(x) \in [\bar{x}_0, x), \ \ \forall x \in (\bar{x}_0, \sqrt{\mu}). \tag{33}$$

(III.2) A consequence of Exp(32, 33) is that any point in I will converge to $\bar{x}_0$ with even steps of gradient descent. For simplicity, we provide the proof for $x \in [x_s, \bar{x}_0)$.

Denote $a_0 \in [x_s, \bar{x}_0)$ and $a_n \coloneqq F_\eta^2(a_{n-1}), n \geq 1$. The series $\{a_i\}_{i \geq 0}$ satisfies

$$\bar{x}_0 \geq a_{n+1} > a_n > a_0. \tag{34}$$

Since the series is bounded and strictly increasing, it is converging. Assume it is converging to a. If $a < \bar{x}_0$, then

$$\bar{x}_0 \geq F_\eta^2(a) > a > F_\eta^2(a_n).$$

Since $F_\eta^2(\cdot)$ is continuous, so $\exists\, \delta > 0$, such that, when $|x - a| < \delta$, we have

$$|F_\eta^2(x) - F_\eta^2(a)| < F_\eta^2(a) - a. \tag{35}$$

Since a is the limit, so $\exists\, N > 0$, such that, when $n > N$, $0 < a - F_\eta^2(a_n) < \delta$. So, combining with Exp(35), we have

$$|F_\eta^2(F_\eta^2(a_n)) - F_\eta^2(a)| < F_\eta^2(a) - a.$$

But LHS $= F_\eta^2(a) - a_{n+2} > F_\eta^2(a) - a$, so we reach a contradiction.

Hence, we have $\{a_i\}$ converges to $\bar{x}_0$.

(III.3) Obviously, any initialization in $(0, \sqrt{\mu})$ will have gradient descent run into (i) the interval I, or (ii) the interval on the right of $\sqrt{\mu}$, *i.e.*, $(\sqrt{\mu}, \infty)$. The first case is exactly our target.

Now consider the second case. From the definition of x_s in part III.1, we know $F_\eta(x_s) = \max_{x\in[0,\sqrt{\mu}]} F_\eta(x)$. So it is fair to say this case is $x_n \in (\sqrt{\mu}, F_\eta(x_s)]$. Then the next step will go into the interval I, because

$$F_\eta(x_n) \geq F_\eta(F_\eta(x_s)) = F_\eta^2(x_s) > x_s,$$

where the first inequality is from the decreasing property of $F_\eta(\cdot)$ and the second inequality is due to $F_\eta^2(x) > x$ on $x \in [x_s, \bar{x}_0)$. □

E Proof of Theorem 3

Theorem 7 (Restatement of Theorem 3)**.** *For $f(x, y) = \frac{1}{2}(xy - \mu)^2$, consider GD with learning rate $\eta = K \cdot \frac{1}{\mu}$. Assume both x and y are always positive during the whole process $\{x_i, y_i\}_{i\geq 0}$. In this process, denote a series of all points with $xy > \mu$ as $\mathcal{P} = \{(x_i, y_i) | x_i y_i > \mu\}$. Then $|x - y|$ decays to 0 in $\mathcal{P}$, for any $1 < K < 1.5$.*

Proof. Consider the current step is at (x_t, y_t) with $x_t y_t > \mu$. After two steps of gradient descent, we have

$$x_{t+1} = x_t + \eta(\mu - x_t y_t) y_t \tag{36}$$
$$y_{t+1} = y_t + \eta(\mu - x_t y_t) x_t \tag{37}$$
$$x_{t+2} = x_{t+1} + \eta(\mu - x_{t+1} y_{t+1}) y_{t+1} \tag{38}$$
$$y_{t+2} = y_{t+1} + \eta(\mu - x_{t+1} y_{t+1}) x_{t+1}, \tag{39}$$

with which we have the difference evolve as

$$y_{t+1} - x_{t+1} = (y_t - x_t)(1 - \eta(\mu - x_t y_t)) \tag{40}$$
$$y_{t+2} - x_{t+2} = (y_{t+1} - x_{t+1})(1 - \eta(\mu - x_{t+1} y_{t+1})). \tag{41}$$

Meanwhile, we have

$$\begin{aligned} x_{t+1} y_{t+1} &= x_t y_t + \eta(\mu - x_t y_t)(x_t^2 + y_t^2) + \eta^2 (\mu - x_t y_t)^2 x_t y_t \\ &= x_t y_t (1 + \eta(\mu - x_t y_t))^2 + \eta(\mu - x_t y_t)(x_t - y_t)^2 \end{aligned} \tag{42}$$

Note that the second term in Eq(42) vanishes when x and y are balanced. When they are not balanced, if $x_t y_t > \mu$, it holds $x_{t+1} y_{t+1} < x_t y_t (1 + \eta(\mu - x_t y_t))^2$. Incorporating this inequality into Eq(40, 41) and assuming $y_t - x_t > 0$, it holds

$$y_{t+2} - x_{t+2} < (y_t - x_t)(1 - \eta(\mu - x_t y_t)) \left(1 - \eta\left(\mu - x_t y_t (1 + \eta(\mu - x_t y_t))^2\right)\right). \tag{43}$$

To show that $|x - y|$ is decaying as in the theorem, we are to show

1. $y_{t+2} - x_{t+2} < y_t - x_t$

2. $y_{t+2} - x_{t+2} > -(y_t - x_t)$

Note that, although $x_t y_t > \mu$, it is not sure to have $x_{t+2}y_{t+2} > \mu$. However, for any $0 < x_i y_i < \mu$ and $K < 2$, we have

$$\frac{|x_{i+1} - y_{i+1}|}{|x_i - y_i|} = |1 - \eta(\mu - x_i y_i)| < 1, \tag{44}$$

which is saying $|x - y|$ decays until it reaches $xy > \mu$. So it is enough to prove the above two inequalities, whether or not $x_{t+2}y_{t+2} > \mu$.

Part I. To show $y_{t+2} - x_{t+2} < y_t - x_t$

Since we wish to have $y_{t+2} - x_{t+2} < y_t - x_t$, it is sufficient to require

$$(1 - \eta(\mu - x_t y_t))\left(1 - \eta\left(\mu - x_t y_t (1 + \eta(\mu - x_t y_t))^2\right)\right) < 1. \tag{45}$$

Since we assume $x_{t+1}, y_{t+1} > 0$, Eq (36, 37) tells $\eta(\mu - x_t y_t) > -\min\{\frac{x_t}{y_t}, \frac{y_t}{x_t}\}$, which is equivalent to $1 - \eta(\mu - x_t y_t) < 1 + \min\{\frac{x_t}{y_t}, \frac{y_t}{x_t}\}$.

(I.1) If $\eta(\mu - x_{t+1}y_{t+1}) \geq \frac{1}{2}$

Then we have $1 - \eta(\mu - x_{t+1}y_{t+1}) \leq \frac{1}{2}$. As a result,

$$\frac{y_{t+2} - x_{t+2}}{y_t - x_t} = (1 - \eta(\mu - x_t y_t))(1 - \eta(\mu - x_{t+1}y_{t+1})) < \left(1 + \min\{\frac{x_t}{y_t}, \frac{y_t}{x_t}\}\right) \times \frac{1}{2} \tag{46}$$

$$= \frac{1}{2} + \frac{1}{2}\min\{\frac{x_t}{y_t}, \frac{y_t}{x_t}\} \tag{47}$$

(I.2) If $\eta(\mu - x_{t+1}y_{t+1}) < \frac{1}{2}$ **and** $x_{t+1}y_{t+1} \leq x_s^2 = \frac{1}{3}\left(\mu + \frac{1}{\eta}\right)$

The second condition reveals

$$\frac{y_{t+2} - x_{t+2}}{y_{t+1} - x_{t+1}} = 1 - \eta(\mu - x_{t+1}y_{t+1}) \leq 1 - \eta\left(\mu - \frac{1}{3}\left(\mu + \frac{1}{\eta}\right)\right)$$
$$= \frac{4}{3} - \frac{2}{3}K. \tag{48}$$

The first condition is equivalent to $x_{t+1}y_{t+1} > \mu - \frac{1}{2\eta}$. Since the second term in Eq(42) is negative, we have

$$x_t y_t (1 + \eta(\mu - x_t y_t))^2 > \mu - \frac{1}{2\eta}, \tag{49}$$

with which we would like to find an upper bound of $x_t y_t$.

Denoting $b = x_t y_t$, consider a function $q(b) = b(1 + \eta(\mu - b))^2$. Obviously $q(\mu) = \mu$. Its derivative is $q'(b) = (1 + \eta\mu - \eta b)(1 + \eta\mu - 3\eta b) < 0$ on the domain of our interest. If we can show an (negative) upper bound for the derivative as $q'(b) < -1$ on a proper domain, then it is fair to say that, from Exp(49), $x_t y_t < \mu + \frac{1}{2\eta}$. Then we have

$$\frac{y_{t+1} - x_{t+1}}{y_t - x_t} = 1 - \eta(\mu - x_t y_t) < 1 - \eta\left(\mu - \left(\mu - \frac{1}{2\eta}\right)\right) = \frac{3}{2}. \tag{50}$$

Then, combining Exp(50, 48), it tells

$$\frac{y_{t+2} - x_{t+2}}{y_t - x_t} < 2 - K. \tag{51}$$

The remaining is to show $q'(b) < -1$ on a proper domain. We have $q'(b) = (1 + \eta\mu - 2\eta b)^2 - (\eta b)^2$, which is equal to $1 - 2\eta\mu < -1$ when $b = \mu$. Meanwhile, the derivative of $q'(b)$ is $q''(b) =$

$-2\eta(\eta b + (1+\eta\mu - 2\eta b)) = -2\eta(1+\eta\mu-\eta b)$, which is negative when $b < \mu + \frac{1}{\eta}$. As a result, it always holds $q'(b) < -1$ when $b < \mu + \frac{1}{\eta}$.

(I.3) If $x_{t+1}y_{t+1} \geq x_s^2$

Denoting again $b = x_t y_t$, the above inequality in is saying, with $b > \mu$,

$$p(b) = (1-\eta(\mu-b))\left(1-\eta\left(\mu - b(1+\eta(\mu-b))^2\right)\right) < 1. \tag{52}$$

After expanding $p(\cdot)$, we have

$$p(b) - 1 = \eta(\mu-b)\left(-2+\eta(\mu-b)+2\eta b - \eta^2 b(\mu-b) - \eta^3 b(\mu-b)^2\right).$$

Apparently $p(\mu) = 1$. So it is necessary to investigate whether $p'(b) < 0$ on $b > \mu$, as

$$p'(b) = 2 - 2\eta b + (\mu - b)\left(\eta^2(1+\eta(\mu-b))(-\mu+3b) + \eta^3 b(\mu-b)\right).$$

Since $\eta b > 1$ and $b > \mu$, it is enough to require

$$(1+\eta(\mu-b))(-\mu+3b) + \eta b(\mu-b) > 0$$
$$(1+\eta(\mu-b))(-\mu+b) + \eta b(\mu-b) + 2b(1+\eta(\mu-b)) > 0.$$

It suffices to show

$$\eta(\mu-b) + 2(1+\eta(\mu-b)) = 2 + 3\eta(\mu-b) > 0. \tag{53}$$

Since $x_{t+1}y_{t+1} \geq x_s^2 = \frac{1}{3}\left(\mu + \frac{1}{\eta}\right)$, it holds

$$b(1+\eta(\mu-b))^2 \geq \frac{1}{3}\left(\mu + \frac{1}{\eta}\right)$$
$$2 + 3\eta(\mu-b) \geq \sqrt{\frac{3\left(\mu+\frac{1}{\eta}\right)}{b}} - 1 > 0,$$

where the last inequality holds because: if $b \geq 3\left(\mu + \frac{1}{\eta}\right)$, then $1+\eta(\mu-b) \leq -2\eta\mu - 2 < 0$, which contradicts with the assumption that both x_{t+1}, y_{t+1} are positive. As a result, the above argument gives

$$\frac{y_{t+2}-x_{t+2}}{y_t - x_t} < p(b) < 1 - 2(K-1)(b-\mu). \tag{54}$$

Part II. To show $y_{t+2} - x_{t+2} > -(y_t - x_t)$

Since $x_t y_t > \mu$, we have $1 - \eta(\mu - x_t y_t) > 1$. Combining with $1 - \eta(\mu - x_t y_t) < 2$, it holds

$$\frac{y_{t+1}-x_{t+1}}{y_t - x_t} = 1 - \eta(\mu - x_t y_t) \in (1, 2).$$

So the remaining is to have $\frac{y_{t+2}-x_{t+2}}{y_{t+1}-x_{t+1}} > -0.5$. Actually it is $1-\eta(\mu - x_{t+1}y_{t+1}) \geq 1 - \eta\mu = 1 - K$. Therefore, we have

$$\frac{y_{t+2}-x_{t+2}}{y_t - x_t} > -1 + (3-2K), \tag{55}$$

as required.

Part III. To show $y_t - x_t$ **converges to 0**

From Exp (47, 51, 54, 55), we have for points in $\mathcal{P}$, $|y-x|$ is a monotone strictly decreasing sequence lower bounded by 0. Hence it is convergent. Actually it converges to 0. If not, assuming it converges to $\epsilon > 0$, the next point will have the difference as $\tilde{\epsilon} < \epsilon$ as well as all following points. Hence, the contradiction gives the convergence to 0. □

F PROOF OF LEMMA 3

Lemma 6 (Restatement of Lemma 3)**.** *In the setting of Theorem 3, denote the initialization as* $m = \frac{|y_0 - x_0|}{\sqrt{\mu}}$ *and* $x_0 y_0 > \mu$*. Then, during the whole process, both* x *and* y *will always stay positive, denoting* $p = \frac{4}{\left(m+\sqrt{m^2+4}\right)^2}$ *and* $q = (1+p)^2$*, if*

$$\max\left\{\eta(x_0y_0-\mu), \frac{4}{27}\left(1+K\right)^3 + \left(\frac{2}{3}K^2 - \frac{1}{3}K + \frac{qK^2}{2(K+1)}m^2\right)qm^2 - K\right\} < p.$$

Proof. Considering $x_t y_t > \mu$, one step of gradient descent returns

$$\begin{aligned} x_{t+1} &= x_t + \eta(\mu - x_t y_t) y_t \\ y_{t+1} &= y_t + \eta(\mu - x_t y_t) x_t. \end{aligned}$$

To have both $x_{t+1} > 0, y_{t+1} > 0$, it suffices to have

$$\eta(x_t y_t - \mu) < \min\left\{\frac{y_t}{x_t}, \frac{x_t}{y_t}\right\}. \tag{56}$$

This inequality will be the main target we need to resolve in this proof.

First, we are to show

$$\min\left\{\frac{y_0}{x_0}, \frac{x_0}{y_0}\right\} > \frac{4}{\left(m+\sqrt{m^2+4}\right)^2}.$$

With the difference fixed as $m = (y_0 - x_0)/\sqrt{\mu}$, assuming $y_0 > x_0$, we have $m/y_0 = (1 - x_0/y_0)/\sqrt{\mu}$. if $x_0 y_0$ increases, both x_0 and y_0 increase then m/y_0 decreases, which means x_0/y_0 increases. As a result, we have

$$\min\left\{\frac{y_0}{x_0}, \frac{x_0}{y_0}\right\} > \min\left\{\frac{y_0}{x_0}, \frac{x_0}{y_0}\right\}\Bigg|_{x_0y_0=\mu} = \frac{4}{\left(m+\sqrt{m^2+4}\right)^2}.$$

Therefore, at initialization, to have positive x_1 and y_1, it is enough to require

$$\eta(x_0y_0 - \mu) < \frac{4}{\left(m+\sqrt{m^2+4}\right)^2} \triangleq r.$$

From Theorem 3, it is guaranteed that $|x_t - y_t| < |x_0 - y_0|$ with $t \geq 2$ until it reaches $x_t y_t > \mu$, with which r is still a good lower bound for $\min\{y_t/x_t, x_t/y_t\}$. So what remains to show is it satisfies $\eta(x_t y_t - \mu) < r$ for the next first time $x_t y_t > \mu$. If this holds, we can always iteratively show, for any $x_t y_t > \mu$ along gradient descent,

$$\eta(x_t y_t - \mu) < r < \min\left\{\frac{y_t}{x_t}, \frac{x_t}{y_t}\right\}.$$

Note that r itself is independent of $x_t y_t$ and all the history, so it is ideal to compute a uniform upper bound of $\eta(x_t y_t - \mu)$ with any pair of (x_{t-1}, y_{t-1}) satisfying $x_{t-1}y_{t-1} < \mu$. Actually it is possible, since we have $|x_{t-1} - y_{t-1}|$ bounded as in Theorem 3.

Assume $x_i y_i > \mu$ and it satisfies the condition of $\eta(x_i y_i - \mu) < r$ and $|x_i - y_i| < |x_0 - y_0|$. As in (40), we have

$$\frac{x_{i+1} - y_{i+1}}{x_i - y_i} = 1 - \eta\left(\mu - x_i y_i\right) \in (1, 1+r). \tag{57}$$

Hence, it suffices to get the maximum value of $g(z)$, with $z \in (0, \mu)$, as

$$g(z) = z\left(1 + \eta(\mu - z)\right)^2 + \eta(\mu - z)(1+r)^2(x_0 - y_0)^2, \tag{58}$$

which is from (42). Denote $\bar{z} = \text{argmax } g(z)$. Obviously $\bar{z} < \frac{1}{3}(\mu + \frac{1}{\eta}) \triangleq z_b$, because the first term of $g(z)$ achieves maximum at $z = \frac{1}{3}(\mu + \frac{1}{\eta})$ and the second term is in a decreasing manner with z. Then let's take the derivative of $g(z)$ as

$$\begin{aligned} g'(z) &= (1 + \eta(\mu - z))\,(1 + \eta\mu - 3\eta z) - \eta(1+r)^2(x_0 - y_0)^2 \\ &= (1 + \eta(\mu - z))\left(1 + \eta\mu - 3\eta z - \frac{\eta(1+r)^2(x_0 - y_0)^2}{1 + \eta(\mu - z)}\right), \end{aligned}$$

where the first term is always positive, so we have

$$1 + \eta\mu - 3\eta\bar{z} - \frac{\eta(1+r)^2(x_0 - y_0)^2}{1 + \eta(\mu - \bar{z})} = 0, \tag{59}$$

which means

$$\begin{aligned} \bar{z} &= \frac{1}{3\eta}\left(1 + \eta\mu - \frac{\eta(1+r)^2(x_0 - y_0)^2}{1 + \eta(\mu - \bar{z})}\right) && (60) \\ &> \frac{1}{3\eta}\left(1 + \eta\mu - \frac{\eta(1+r)^2(x_0 - y_0)^2}{1 + \eta(\mu - \frac{1}{3}(\mu + \frac{1}{\eta}))}\right) && (61) \\ &= \frac{1}{3}\left(\mu + \frac{1}{\eta} - \frac{3(1+r)^2}{2(\eta+1)}(x_0 - y_0)^2\right) && (62) \\ &\triangleq z_s, && (63) \end{aligned}$$

where the inequality is from $\bar{z} < \frac{1}{3}(\mu + \frac{1}{\eta})$. As a result, it is safe to say

$$\begin{aligned} g(z) &\le z\,(1 + \eta(\mu - z))^2\Big|_{z=z_b} + \eta(\mu - z)(1+r)^2(x_0 - y_0)^2\Big|_{z=z_s} && (64) \\ &= \frac{4}{27}(1 + \eta\mu)^3 \cdot \frac{1}{\eta} + \eta(1+r)^2\left(\frac{2}{3}\mu - \frac{1}{3\eta} + \frac{2}{\eta\mu + 1}(x_0 - y_0)^2\right)(x_0 - y_0)^2, && (65) \end{aligned}$$

with which we are able to compute $\max\ \eta(g(z) - \mu)$, which is exactly the final result. □

G PROOF OF THEOREM 4

Theorem 8 (Restatement of Theorem 4)**.** *In the above setting, consider a teacher neuron $\tilde{w} = [1, 0]$ and set the learning rate $K \in (1, 1.1]$. Initialize the student as $\left\|w^{(0)}\right\| = v^{(0)} \triangleq \epsilon \in (0, 0.10]$ and $\langle w^{(0)}, \tilde{w}\rangle \ge 0$. Then, for $t \ge T_1 + 4$, $w_y^{(t)}$ decays to 0 with a linear rate of 0.97 at most, starting from a value < 0.1, where*

$$T_1 \le \left\lceil \log_{2.56} \frac{1.35}{\pi\beta^2} \right\rceil, \qquad \beta = \left(1 + \frac{1.1}{\pi}\right)\epsilon.$$

Proof. We restate the update rules as

$$\Delta v^{(t)} := v^{(t+1)} - v^{(t)} = K w_x^{(t)} \left[(-v^{(t)} w_x^{(t)} + 1) - v^{(t)} w_y^{(t)} \frac{w_y^{(t)}}{w_x^{(t)}} - \frac{1}{\pi} \left(\arctan\left(\frac{w_y^{(t)}}{w_x^{(t)}} \right) - \frac{w_y^{(t)}}{w_x^{(t)}} \right) \right],$$

$$= K w_x^{(t)} \left[(-v^{(t)} w_x^{(t)} + 1) - \frac{1}{\pi} \left(\arctan\left(\frac{w_y^{(t)}}{w_x^{(t)}} \right) - \frac{w_x^{(t)} w_y^{(t)}}{\left\| w^{(t)} \right\|^2} \right) \right]$$

$$+ K \frac{(w_y^{(t)})^2}{v^{(t)}} \left(-(v^{(t)})^2 + \frac{v^{(t)} w_y^{(t)}}{\pi \left\| w^{(t)} \right\|^2} \right) \tag{66}$$

$$\Delta w_x^{(t)} := w_x^{(t+1)} - w_x^{(t)} = K v^{(t)} \left[(-v^{(t)} w_x^{(t)} + 1) - \frac{1}{\pi} \left(\arctan\left(\frac{w_y^{(t)}}{w_x^{(t)}} \right) - \frac{w_x^{(t)} w_y^{(t)}}{\left\| w^{(t)} \right\|^2} \right) \right], \tag{67}$$

$$\Delta w_y^{(t)} = w_y^{(t)} \cdot K \left(-(v^{(t)})^2 + \frac{v^{(t)} w_y^{(t)}}{\pi \left\| w^{(t)} \right\|^2} \right), \tag{68}$$

$$w_y^{(t+1)} = \left| w_y^{(t)} + \Delta w_y^{(t)} \right|. \tag{69}$$

For simplicity, we will omit all superscripts of time t unless clarification is necessary. From (69), if the target is to show w_y decaying with a linear rate, it suffices to bound the factor term in (68) (by a considerable margin) as

$$-2 < K \left(-v^2 + \frac{v w_y}{\pi \left\| w \right\|^2} \right) < 0. \tag{70}$$

The technical part is the second inequality of (70). If $v, w_x > 0$, it is equivalent to

$$v w_x > \frac{w_x w_y}{\pi \left\| w \right\|^2} = \frac{w_x w_y}{\pi (w_x^2 + w_y^2)},$$

where the RHS is smaller than or equal to $\frac{1}{2\pi}$. Hence, $\frac{1}{2\pi}$ is a special threshold with which we will frequently compare $v w_x$. Another important variable to control is $v - w_x$ that reveals how the two layers are balanced. If it is too large, for the iteration $v^{(t+1)} w_x^{(t+1)}$ may explode as shown in the 2-D case.

The main idea of our proof is that

- Stage 1 with $v w_x \leq \frac{w_x w_y}{\pi \|w\|^2}$: in this stage, w_y grows but it grows in a smaller rate than that of v and w_x. Therefore, since we have an upper bound for $v w_x$ to stay in this stage, we are able to compute the upper bound of #iterations to finish this stage, which is T_1 in the theorem. At the end of this stage, both of $v - w_x$ and w_y are bounded under our assumption of initialization.
- Stage 2 with $v w_x > \frac{w_x w_y}{\pi \|w\|^2}$: in this stage, w_y decreases. Since our range of a large learning rate is relatively narrow $(1 < K \leq 1.1)$, we are able to compute bounds of $v w_x, v - w_x$ and w_y. After eight iterations, it falls into (and stays in) a bounded basin of these three terms, in which w_y decays at least in a linear rate.

Stage 1.

We are to show that, in the last iteration of this stage, there are three facts: 1) $v w_x \leq \frac{1}{2\pi}$, 2) $v - w_x \in [-0.017, 0.17]$, and 3) $w_y \leq 0.44$.

At initialization, we assume $v^{(0)} = \left\| w^{(0)} \right\|$. Denote $\alpha_0 = \arctan(w_y^{(0)} / w_x^{(0)}) \in [0, \pi/2]$. So for next iteration we have

$$w_y^{(1)} = v^{(0)} \left(1 + K \left(-(v^{(0)})^2 + \frac{1}{\pi} \sin \alpha_0 \right) \right), \tag{71}$$

$$w_x^{(1)} = v^{(0)} \left[\cos \alpha_0 + K \left(1 - (v^{(0)})^2 \cos \alpha_0 + \frac{\cos \alpha_0 \sin \alpha_0 - \alpha_0}{\pi} \right) \right]. \tag{72}$$

Apparently $w_y^{(1)}$ increases with α_0 increasing. And

$$\begin{aligned}\partial_{\alpha_0} w_x^{(1)} &= v^{(0)}\left[-\sin\alpha_0 + K\left((v^{(0)})^2\sin\alpha_0 + \frac{-\sin^2\alpha_0+\cos^2\alpha_0-1}{\pi}\right)\right]\\ &= v^{(0)}\left[-\sin\alpha_0 + K\left(\left((v^{(0)})^2 - \frac{\sin\alpha_0}{\pi}\right)\sin\alpha_0 + \frac{-\sin^2\alpha_0}{\pi}\right)\right].\end{aligned}$$

Since in stage 1 it holds $\Delta w_y > 0$ which means $-(v^{(0)})^2 + \frac{1}{\pi}\sin\alpha_0 > 0$ in (71). So it follows $\partial_{\alpha_0} w_x^{(1)} \leq 0$. Combining the above arguments, we have

$$\begin{aligned}&w_x^{(1)} \geq w_x^{(1)}|_{\alpha_0=\frac{\pi}{2}} = \frac{K}{2}v^{(0)},\\ &w_y^{(1)} \leq w_y^{(1)}|_{\alpha_0=\frac{\pi}{2}} = \left(1+\frac{K}{\pi} - K(v^{(0)})^2\right)v^{(0)} \leq \left(1+\frac{K}{\pi}\right)v^{(0)},\\ &\frac{w_y^{(1)}}{w_x^{(1)}} \leq \frac{2+\frac{2K}{\pi}}{K} \leq 2.7.\end{aligned}$$

Regarding $\frac{v}{w_y}$, it has $v^{(0)} \geq w_y^{(0)}$ at initialization due to $v^{(0)} = \left\|w^{(0)}\right\|$. From (66, 67, 68), we have $v\Delta v = w_x\Delta w_x + w_y\Delta w_y$. So it holds $v\Delta v \geq y\Delta y$. Meanwhile, $\frac{\Delta w_y}{v} = K(-vw_y + \frac{w_y^2}{\pi\|w\|^2}) \in [0, \frac{K}{\pi}]$. From Lemma 7, given $v^{(t)} \geq w_y^{(t)}$ and $\frac{\Delta w_y}{v} \in [0,1]$ for any t in this stage, it always holds $v^{(t+1)} \geq w_y^{(t+1)}$.

Therefore, it is fair to say

$$\frac{v^{(1)}w_x^{(1)}}{(w_y^{(1)})^2} \geq \frac{1}{2.7}.$$

Additionally, to bound the term $vw_y/\left\|w\right\|^2$ in Δw_y, we would like to show it always has $vw_y \leq \left\|w\right\|^2$. At initialization, it naturally holds. Then, for the every next iteration, given it holds in the last iteration, we have

$$\begin{aligned}&(v+\Delta v)(w_y+\Delta w_y) - [(w_x+\Delta w_x)^2 + (w_y+\Delta w_y)^2]\\ &= (v + \frac{w_x\Delta w_x + w_y\Delta w_y}{v})(w_y+\Delta w_y) - [(w_x+\Delta w_x)^2 + (w_y+\Delta w_y)^2]\\ &= vw_y + v\Delta w_y + w_x\Delta w_x(\frac{w_y}{v} + \frac{\Delta w_y}{v}) + (w_y\Delta w_y + (\Delta w_y)^2)\frac{w_y}{v} - [(w_x+\Delta w_x)^2 + (w_y+\Delta w_y)^2]\\ &\leq vw_y + v\Delta w_y + w_y\Delta w_y\frac{w_y}{v} - (w_x^2 + w_y^2 + 2w_y\Delta w_y + (\Delta w_x)^2)\\ &\leq v\Delta w_y + w_y\Delta w_y\frac{w_y}{v} - 2w_y\Delta w_y - (\Delta w_x)^2\\ &= v\Delta w_y(1-\frac{w_y}{v})^2 - (\Delta w_x)^2\\ &\leq v\Delta w_y - (\Delta w_x)^2\end{aligned}$$

where the first equality uses $v\Delta v = w_x\Delta w_x + w_y\Delta w_y$, the first inequality uses the proven $v \geq w_y$ and $v \geq \Delta w_y$, the second inequality uses the assumption $vw_y \leq \left\|w\right\|^2$. Now we are to show $v\Delta w_y - (\Delta w_x)^2 \leq 0$. We have

$$\begin{aligned}v\Delta w_y - (\Delta w_x)^2 &\leq Kv^2\frac{w_y^2}{\pi\left\|w\right\|^2} - K^2v^2\left(1-\frac{1}{2\pi}-\gamma^{(t)}\right)^2,\\ \gamma^{(t)} &= \frac{1}{\pi}\left(\arctan\left(\frac{w_y^{(t)}}{w_x^{(t)}}\right) - \frac{w_x^{(t)}w_y^{(t)}}{\left\|w^{(t)}\right\|^2}\right).\end{aligned}$$

Since we have proven $w_y^{(1)}/w_x^{(1)} \leq 2.7$, it is easy to check that

$$\frac{1}{\pi\left(1+(\frac{w_x^{(1)}}{w_y^{(1)}})^2\right)} \leq (1-\frac{1}{2\pi}-\gamma^{(1)})^2.$$

As a result, $v\Delta w_y - (\Delta w_x)^2 \leq 0$ at time 1. Furthermore, by checking each term, $v\Delta w_y - (\Delta w_x)^2$ decreases with w_y/w_x decreasing. We will soon show that w_y/w_x itself decreases, by showing the growth ratio of w_x is larger than that of w_y.

Our target lower bound of the growth ratio of w_x is that

$$\frac{\Delta w_x}{w_x} \geq 1-\frac{1}{\pi}-\gamma, \tag{73}$$

which is larger than the growth ratio of w_y bounded by $\frac{1}{\pi}$ due to $v\Delta w_y < \|w\|^2$. So it suffices to show $Kv/w_x \geq 1$. Assuming $Kv/w_x \geq 1$ for the current step, we need to show $Kv^{(t+1)}/w_x^{(t+1)} \geq 1$ also holds for the next step. Let's denote

$$A^{(t)} = K\left[(-v^{(t)}w_x^{(t)}+1) - \frac{1}{\pi}\left(\arctan\left(\frac{w_y^{(t)}}{w_x^{(t)}}\right) - \frac{w_x^{(t)}w_y^{(t)}}{\left\|w^{(t)}\right\|^2}\right)\right]. \tag{74}$$

Then

$$\begin{aligned}(v+\Delta v) - \frac{1}{K}(w_x+\Delta w_x) &\geq v + Aw_x - \frac{w_x}{K} - \frac{Av}{K} \\ &= (v-\frac{w_x}{K})(1-KA) + v(K-\frac{1}{K})A. \end{aligned} \tag{75}$$

If $KA \leq 1$, since $K > 1$ and $A > 0$, we have (75) as positive, which is what we need. If $KA > 1$, then

$$\begin{aligned}(75) &\geq (v-\frac{w_x}{K})(1-K^2) + v(K-\frac{1}{K})A \\ &= ((-K+A)v+w_x)\,(K-\frac{1}{K}),\end{aligned}$$

where the first inequality is due to $A \leq K$ and the assumption of $Kv^{(t)}/w_x^{(t)} \geq 1$. Then it suffices to show $(-K+A)v + w_x \geq (-K+\frac{1}{K})v + w_x \geq 0$. Note that $-K+1/K \in (-0.2, 0]$ when $K \in (1, 1.1]$. It is easy to verify that $v^{(1)} \leq 5w_x^{(1)}$. Then, for the next step, we need to show $v^{(t+1)} \leq 5w_x^{(t+1)}$ given $v^{(t)} \leq 5w^{(t+1)}$. To prove this, we are to bound $v - w_x$, as

$$\begin{aligned}v^{(t+1)} - w_x^{(t+1)} &= (1-A)(v-w) + K\frac{w_y^2}{v}(-v^2 + \frac{vw_y}{\pi\left\|w\right\|^2}) \\ &\leq 0.4(v-w) + Kw_y\frac{w_y^2}{\pi\left\|w\right\|^2} \leq 0.4(v-w) + \frac{Kw_y}{\pi},\end{aligned} \tag{76}$$

where the first inequality is due to, when $w_y/w_x \leq 2.7$,

$$\begin{aligned}A &= K\left[-v^{(t)}w_x^{(t)} + \frac{1}{\pi}\frac{w_x^{(t)}w_y^{(t)}}{\left\|w^{(t)}\right\|^2}\right] + K\left[1-\frac{1}{\pi}\arctan\left(\frac{w_y^{(t)}}{w_x^{(t)}}\right)\right] \\ &\geq K\left[1-\frac{1}{\pi}\arctan\left(\frac{w_y^{(t)}}{w_x^{(t)}}\right)\right] \geq 0.6.\end{aligned}$$

We will later show that $v^{(t+1)} - w^{(t+1)} \geq -0.1(v^{(t)} - w^{(t)})$. Combining this with (76), it is safe to say

$$v^{(t+1)} - w^{(t+1)} \leq 0.4(v-w) + \frac{Kw_y}{\pi} \leq 0.4\times 4w + \frac{K\times 5w}{\pi} \leq 4w,$$

where the second inequality is due to $v \leq 5w$ and $v \geq w_y$. Since $w^{(t+1)} \geq w^{(t)}$ (due to $A > 0$) in this stage, we have $v^{(t+1)} \leq 5w_x^{(t+1)}$.

Combining the above discussion, we have prove (73). Obviously, when $w_y/w_x \leq 2.7$, RHS of (73) is at least 0.55, larger than $1.1/\pi$, which is the upper bound of the $\Delta w_y/w_y$. As a result, w_y/w_x keeps decreasing.

The next step is to show the growing ratio of vw_x is much larger than that of w_y. From (67, 68), it holds

$$\begin{aligned} v^{(t+1)}w_x^{(t+1)} &= (v+\Delta v)(w_x+\Delta w_x) \geq vw_x + KA(v^2+w_x^2) + K^2A^2vw_x \\ &\geq vw_x(1+A)^2, \end{aligned}$$

where the first inequality is due to $\Delta w_y \geq 0$. It follows $v^{(t+1)}w_x^{(t+1)}/v^{(t)}w_x^{(t)} \geq 1.6^2 = 2.56$.

So far, we have shown the following facts: under the defined initialization at time 0, starting from time 1, we have

1. $vw_x \leq 1/2\pi$.
2. $\Delta w_x/w_x + 1 \geq 1.55$.
3. $\Delta w_y/w_y + 1 \leq 1 + K/\pi$.
4. $w_y/w_x \leq 2.7$ and keeps decreasing.
5. $v^{(t+1)}w_x^{(t+1)}/v^{(t)}w_x^{(t)} \geq 2.56$.
6. $v \geq w_y$.
7. $v\Delta w_y < (\Delta w_x)^2$.

Now we are to use the above facts to bound vw_x, w_y and $v - w_x$ to the end of stage 1.

For vw_x, in previous discussion, we have shown that $vw_x \leq \frac{1}{2\pi}$. Actually, there is another special value

$$\frac{w_x w_y}{\pi(w_x^2+w_y^2)} = 0.104 \text{ when } w_y/w_x = 2.7. \tag{77}$$

This value is slightly larger than $1/4\pi$. Hence, we would like to split the analysis into three parts: in the **first step of stage 2**,

1. $vw_x \geq \frac{1}{2\pi}$.
2. $\frac{1}{4\pi} \leq vw_x < \frac{1}{2\pi}$.
3. $vw_x < \frac{1}{4\pi}$.

Note that, although we are discussing the stage 1 in this section, investigating the lower bound of the first step in stage 2 helps calculate the number of iterations in stage 1. Furthermore, it helps bound several variables in stage 1.

Case (I). If $vw_x \geq \frac{1}{2\pi}$ in first step of stage 2:

Since we have prove $\frac{v^{(1)}w_x^{(1)}}{(w_y^{(1)})^2} \geq 1/2.7$ and $v^{(t+1)}w_x^{(t+1)}/v^{(t)}w_x^{(t)} \geq 2.56$, the number of iterations for vw_x to reach $1/2\pi$ is at most

$$T_1 \leq \left\lceil \log_{2.56} \frac{\frac{1}{2\pi}}{(w_y^{(1)})^2/2.7} \right\rceil. \tag{78}$$

Meanwhile, starting from time 1, the growth ratio of w_y is

$$(w_y + \Delta w_y)/w_y \leq 1 + K(-v^2 + 1/\pi) \leq 1 + 1.1/\pi - (v^{(1)})^2 \leq 1 + 1.1/\pi - (w_y^{(1)})^2, \quad (79)$$

where the first inequality is due to $vw_y \leq \|w\|^2$, the second is due to $K > 1$ and the third is from $v \geq w_y$. Therefore, combining with (78), we can bound w_y in the end of stage 1 as

$$w_y \leq \left(1 + 1.1/\pi - (w_y^{(1)})^2\right)^{\left\lceil \log_{2.56} \frac{\frac{1}{2\pi}}{(w_y^{(1)})^2/2.7} \right\rceil}. \quad (80)$$

Since it initializes as $\left\|w^{(0)}\right\| \leq 0.1$, we have $w_y^{(1)} \leq 0.1(1 + 1.1/\pi) = 0.135$. Then, it can be verified that, when $w_y^{(1)} \in (0, 0.135]$, it holds

$$w_y \leq 0.44. \quad (81)$$

The next is to bound $v - w_x$. Combining the update rules of v and w_x in (66, 67), we have

$$\begin{aligned}\Delta(v - w_x) &\coloneqq (v^{(t+1)} - w_x^{(t+1)}) - (v^{(t)} - w_x^{(t)}) \\ &= K(v - w_x)\left(vw_x - 1 + \frac{\arctan(w_y/w_x) - \frac{w_x w_y}{\|w\|^2}}{\pi}\right) + K\frac{w_y^2}{v}(-v^2 + \frac{vw_y}{\pi\|w\|^2}).\end{aligned} \quad (82)$$

Note that

$$-1 \leq vw_x - 1 + \frac{\arctan(w_y/w_x) - \frac{w_x w_y}{\|w\|^2}}{\pi} \leq -1 + \frac{\arctan(w_y/w_x)}{\pi}, \quad (83)$$

where the left is due to $vw_x > 0$ and , the right is from $\Delta w_y \geq 0$. When $w_y/w_x \leq 2.7$, the RHS follows $-1 + \frac{\arctan(w_y/w_x)}{\pi} \leq -0.6$. So combining both sides tells

$$1 + K\left(vw_x - 1 + \frac{\arctan(w_y/w_x) - \frac{w_x w_y}{\|w\|^2}}{\pi}\right) \in [-K + 1, 0.4] \subset [-0.1, 0.4]. \quad (84)$$

Since $\Delta w_y \geq 0$, we have $0 \leq K\frac{w_y^2}{v}(-v^2 + \frac{vw_y}{\pi\|w\|^2}) \leq \frac{K}{\pi} w_y \frac{w_y^2}{\|w\|^2}$. Note that at initialization $w_x^{(0)} \leq v^{(0)}$. Then it is easy to verify that

$$-0.01 \leq -0.1(v^{(0)} - w^{(0)}) \leq v^{(1)} - w^{(1)} \leq (1 + \frac{K}{\pi} - \frac{K}{2})v^{(0)} \leq 0.082. \quad (85)$$

Because the coefficient on the positive side in (84) is larger than $0.4 > 0.1$, it is appropriate to upper bound the $v - w_x$ as

$$\begin{aligned}v - w_x &\leq \max\left\{0.082, 0.082 \cdot 0.4^T + \sum_{t=1}^{T} 0.4^{t-1}\frac{K}{\pi} w_y^{(t)} \frac{(w_y^{(t)})^2}{\left\|w^{(t)}\right\|^2}\right\} \\ &\leq \max\left\{0.082, 0.082 \cdot 0.4^T + \sum_{t=1}^{T} 0.4^{t-1}\frac{K}{\pi} w_y^{(t)} \frac{1}{1 + \frac{1}{2.7}\left(\frac{1.55}{1+K/\pi}\right)^{2(t-1)}}\right\} \\ &\leq \max\left\{0.082, 0.082 \cdot 0.4^T + \sum_{t=1}^{T} 0.4^{t-1}\frac{1.1 \cdot 4.4}{\pi} \frac{1}{1 + \frac{1}{2.7}\left(\frac{1.55}{1+1.1/\pi}\right)^{2(t-1)}}\right\},\end{aligned}$$

where the second inequality is from the different growth ratios of w_x and w_y. Note that here we take all $T \geq 1$ and pick the largest value of RHS to bound w_y. It turns out

$$v - w_x \leq 0.17. \quad (86)$$

Furthermore, to lower bound $v - w_x$, since obviously $|v - w_x| \leq 0.17$, it follows

$$v - w_x \geq -0.1 \cdot |v - w_x|_{\max} \geq -0.017. \tag{87}$$

Case (II). If $\frac{1}{4\pi} \leq vw_x < \frac{1}{2\pi}$ in first step of stage 2:

Similar to the discussion in Case (I), we are able to compute the number of iterations for vw_x to reach $1/4\pi$. It is at most

$$T_1 \leq \lceil \log_{2.56} \frac{\frac{1}{4\pi}}{(w_y^{(1)})^2/2.7} \rceil. \tag{88}$$

Accordingly, w_y is bounded as

$$w_y \leq \left(1 + 1.1/\pi - (w_y^{(1)})^2\right)^{\lceil \log_{2.56} \frac{\frac{1}{4\pi}}{(w_y^{(1)})^2/2.7} \rceil} \leq 0.37. \tag{89}$$

For simplicity, we just keep the bounds for $v - w_x$ as in Case (I), as

$$v - w_x \in [-0.017, 0.17]. \tag{90}$$

Case (III). If $vw_x < \frac{1}{4\pi}$ in first step of stage 2:

From the condition, we know $vw_x < \frac{1}{4\pi}$ as well in the last step of stage 1. Since $\Delta w_y > 0$ in stage 1, it tells

$$\frac{1}{\pi} \frac{w_x w_y}{\|w\|^2} < vw_x \leq \frac{1}{4\pi}, \tag{91}$$

which means

$$\max\{\frac{w_x}{w_y}, \frac{w_y}{w_x}\} \geq 2 + \sqrt{3}. \tag{92}$$

Since $2 + \sqrt{3} > 2.7$, if $w_y/w_x \geq 2 + \sqrt{3}$, then for time 1, $(v^{(1)}, w_x^{(1)}, w_y^{(1)})$ is already in the stage 2. However, it is not possible because $\left\|w^{(0)}\right\| = v^{(0)} \leq 0.1$, which means $v^{(1)} w_x^{(1)}$ can not reach $\frac{1}{\pi} \frac{2.7}{1+2.7^2}$.

Therefore, the only possible is $\frac{w_x}{w_y} \geq 2 + \sqrt{3}$. In this case, we are able to bound w_y as

$$w_y \leq (2 - \sqrt{3}) w_x \leq (2 - \sqrt{3}) \left(\sqrt{\frac{1}{4\pi} + 0.0085^2} + 0.0085 \right) \leq 0.078, \tag{93}$$

where the second inequality is due to $vw_x \leq \frac{1}{4\pi}$ and $v - w_x \geq -0.017$. Note that here we still use the bound of $v - w_x$ from Case (I), although it is loose somehow but it is enough for our analysis.

We leave the analysis of the bound of number of iterations to the end of this section.

Stage 2.

In the case (I) of stage 1, where the first step in stage 2 is with $vw_x \geq \frac{1}{2\pi}$, it has $v - w_x \in [-0.017, 0.17]$ and $w_y \leq 0.44$. In the case (II), where the first step of stage 2 is with $vw_x \in [\frac{1}{4\pi}, \frac{1}{2\pi}]$, it has $v - w_x \in [-0.017, 0.17]$ and $w_y \leq 0.37$. In the case (III), where the first step of stage 2 is with $vw_x \in [\frac{1}{4\pi}, \frac{1}{2\pi}]$, it has $v - w_x \in [-0.017, 0.17]$ and $w_y \leq 0.078$.

To upper bound vw_x in the first step of stage 2, there are two candidates. One is from the case (I),

$$\begin{aligned}
v^{(t+1)}w_x^{(t+1)} &= vw_x\left(1+K(1-vw_x-\frac{\arctan(\frac{w_y}{w_x})-\frac{w_y/w_x}{1+(w_y/w_x)^2}}{\pi})\right)^2+K\frac{w_xw_y^2}{v}\left(-v^2+\frac{vw_y}{\pi\left\|w\right\|^2}\right)\\
&\quad+K(v-w_x)^2\left(1+K(1-vw_x-\frac{\arctan(\frac{w_y}{w_x})-\frac{w_y/w_x}{1+(w_y/w_x)^2}}{\pi})\right)\\
&\le vw_x\left(1+K(1-vw_x)\right)^2+K\frac{w_xw_y^2}{w_x}\left(-vw_x+\frac{w_xw_y}{\pi\left\|w\right\|^2}\right)\\
&\quad+K(v-w_x)^2\left(1+K(1-vw_x)\right)\\
&\le \frac{1}{2\pi}\left(1+1.1(1-\frac{1}{2\pi})\right)^2+1.1\cdot 0.44^2\left(-\frac{1}{4\pi}+\frac{1}{2\pi}\right)+1.1\cdot 0.17^2\left(1+1.1(1-\frac{1}{2\pi})\right)\\
&\le 0.668, \qquad (94)
\end{aligned}$$

where we use $vw_x \ge 1/4\pi$, $x/(1+x^2)\le 0.5$ for any x.

One is from the case (II),

$$\begin{aligned}
v^{(t+1)}w_x^{(t+1)} &\le vw_x\left(1+K(1-vw_x)\right)^2+K\frac{w_xw_y^2}{w_x}\left(-vw_x+\frac{w_xw_y}{\pi\left\|w\right\|^2}\right)\\
&\quad+K(v-w_x)^2\left(1+K(1-vw_x)\right)\\
&\le \frac{1}{4\pi}\left(1+1.1(1-\frac{1}{4\pi})\right)^2+1.1\cdot 0.37^2\left(\frac{1}{2\pi}\right)+1.1\cdot 0.17^2\left(1+1.1(1-\frac{1}{4\pi})\right)\\
&\le 0.48, \qquad (95)
\end{aligned}$$

where we use $vw_x \le 1/4\pi$, $x/(1+x^2)\le 0.5$ for any x.

Therefore, we can see that, in the first step of stage 2,

$$vw_x \le 0.668. \qquad (96)$$

Next we are going to show how the iteration goes in the stage 2. In Case (I), there are three facts:

1. $w_y \le 0.44$.
2. $v-w_x \in [-0.017, 0.17]$.
3. $vw_x \in [\frac{1}{2\pi}, 0.668]$.

Similarly, in Case (II), there are three facts as well:

1. $w_y \le 0.37$.
2. $v-w_x \in [-0.017, 0.17]$.
3. $vw_x \in [\frac{1}{4\pi}, \frac{1}{2\pi}]$.

The main idea is to find a basin that any iteration with the above properties (*i.e.*, in the interval) will converge to and then stay in. The method is to iteratively compute the ranges of the variables for several steps, thanks to the narrow range of K. Before explicitly computing the ranges, let's write down the computing method, depending on whether or not $vw_x \ge 1$.

Consider any iteration with $vw_x \in [m_1, m_2], v-w_x \in [d_1, d_2], w_y \le e$, we compute the bounds of $v^{(t+1)}w_x^{(t+1)}, v^{(t+1)}-w_x^{(t+1)}, w_y^{(t+1)}$ in the following process (naturally assuming $d_1 < 0 < d_2$)

1. If $m_1 \ge 1$:

(a) Compute $w_x \geq \sqrt{m_1 + (d_2/2)^2} - d_2/2 \triangleq f$.

(b) Compute $\frac{w_y}{w_x} \leq e/f \triangleq g$.

(c) Compute $\frac{\arctan(w_y/w_x) - \frac{w_x w_y}{\|w\|^2}}{\pi} \leq \frac{\arctan(g) - g/(1+g^2)}{\pi} \triangleq h$.

(d) Compute $v^{(t+1)} w_x^{(t+1)} \geq m_2(1 + 1.1(1 - m_2 - h))^2 + 1.1(1 - m_2 - h) \max\{|d_1|, |d_2|\}^2 - 1.1e^2 m_2$. This is from

$$\begin{aligned} v^{(t+1)} w_x^{(t+1)} &\geq v w_x \left(1 + K(1 - v w_x - h)\right)^2 + K \frac{w_x w_y^2}{v} \left(-v^2 + \frac{v w_y}{\pi \|w\|^2}\right) \\ &\quad + K(v - w_x)^2 \left(1 + K(1 - v w_x - h)\right) \\ &\geq v w_x \left(1 + K(1 - v w_x - h)\right)^2 - K w_y^2 \cdot v w_x \\ &\quad + K(v - w_x)^2 \left(1 + K(1 - v w_x - h)\right). \end{aligned}$$

(e) Compute $v^{(t+1)} w_x^{(t+1)} \leq m_1(1 + 1.0(1 - m_1))^2$. This is due to $x(1 + K(1 - x))^2$ decreases with x increasing when $x \geq 1$.

(f) Compute $v^{(t+1)} - w_x^{(t+1)} \in [d_1(1 + 1.1(m_2 - 1 + h) - 1.1e^2 \cdot (\sqrt{m_2 + (d_2/2)^2} + d_2/2)), d_2(1 + 1.1(m_2 - 1 + h))]$. This is due to

$$\Delta v - \Delta w_x = K(v - w_x) \left(v w_x - 1 + \frac{1}{\pi}(\arctan(\alpha) - \frac{w_x w_y}{\|w\|^2})\right) + K \frac{w_y^2}{v} \left(-v^2 + \frac{v w_y}{\pi \|w\|^2}\right),$$

where $v w_x \geq 1$, the last term is between $-K v w_y^2$ and 0.

(g) Compute $w_y^{(t+1)} \leq e \cdot \max\{|j_1|, |j_2|\}$, where

$$j_1 = 1 + 1.1 \frac{\sqrt{m_1 + (d_2/2)^2} + d_2/2}{\sqrt{m_1 + (d_2/2)^2} - d_2/2} \cdot (-m_2), \tag{97}$$

$$j_2 = 1 + 1.0 \frac{\sqrt{m_1 + (d_1/2)^2} - d_1/2}{\sqrt{m_1 + (d_1/2)^2} + d_1/2} \cdot (-m_1 + \frac{1}{2\pi}). \tag{98}$$

This is due to

$$\frac{\Delta w_y}{w_y} = K \frac{v}{w_x} (-v w_x + \frac{1}{\pi} \frac{w_x w_y}{\|w\|^2}),$$

then we would like to have the smallest value as $j_1 - 1$ and the largest value as $j_2 - 1$. Since w_y is always non-negative, taking the maximum absolute value gives the upper bound.

2. If $m_2 < 1$:

(a) Compute $w_x \geq \sqrt{m_1 + (d_2/2)^2} - d_2/2 \triangleq f$.

(b) Compute $\frac{w_y}{w_x} \leq e/f \triangleq g$.

(c) Compute $\frac{\arctan(w_y/w_x) - \frac{w_x w_y}{\|w\|^2}}{\pi} \leq \frac{\arctan(g) - g/(1+g^2)}{\pi} \triangleq h$.

(d) Compute $v^{(t+1)} w_x^{(t+1)} \geq \min_{x \in [m_1, m_2]} x(1 + 1.0(1 - x - h))^2 - 1.1e^2 x$. Compared with the case of $m_1 \geq 1$, we drop the term $1.1(1 - m_2 - h) \max\{|d_1|, |d_2|\}^2$ because it is possible to have $v - w_x = 0$ in some iterations.

(e) Compute $v^{(t+1)} w_x^{(t+1)} \leq \max_{x \in [m_1, m_2]} x(1 + 1.1(1 - x))^2 + 1.1(1 - x) \max\{|d_1|, |d_2|\}^2$. Compared with the case of $m_1 \geq 1$, we add a term depending on the $|v - w_x|_{\max}$ because it enlarges $v w_x$ in the in-balanced case.

(f) Compute $v^{(t+1)} - w_x^{(t+1)} \in [d_1(1 + 1.1(m_2 - 1 + h) - 1.1e^2 \cdot (\sqrt{m_2 + (d_2/2)^2} + d_2/2)), d_2(1 + 1.1(m_2 - 1 + h))]$. In fact, a rigorous left bound should include more terms to select a minimum from. Here it is simple because it keeps $1 + K(m_1 - 1) \geq 0$ in the following computing, so we do not need to worry about the flipping sign of d_1 and d_2.

(g) Compute $w_y^{(t+1)} \leq e \cdot \max\{|j_1|, |j_2|\}$, where j_1, j_2 are the same with those in the case of $m_1 \geq 1$.

Therefore, with the above process, we are able to brutally compute the ranges of $v^{(t+1)}w_x^{(t+1)}, v^{(t+1)} - w_x^{(t+1)}, w_y^{(t+1)}$ from the current ranges. Note that this process plays a role of building a mapping from one interval to another interval, which covers all points from the source interval. However, it is loose to some extent because gradient descent is a mapping from a point to another point. The advantage of such a loose method is feasibility of obtaining bounds while losing tightness. To achieve tightness, later we will also include some wisdom in a point-to-point style.

Also note that, a nice way to combine tightness and efficiency in this method is to split and to merge intervals when necessary.

For Case (I):

Now we are to compute the ranges starting from the interval where $I = \{w_y \leq 0.44, v - w_x \in [-0.017, 0.17], vw_x \in [\frac{1}{2\pi}, 0.668]\}$. First, we split it into three intervals:

1. $I_1 = \{w_y \leq 0.44, v - w_x \in [-0.017, 0.17], vw_x \in [0.213, 0.4]\}$.
2. $I_2 = \{w_y \leq 0.44, v - w_x \in [-0.017, 0.17], vw_x \in [0.4, 0.668]\}$.
3. $I_{30} = \{w_y \leq 0.44, v - w_x \in [-0.017, 0.17], vw_x \in [\frac{1}{2\pi}, 0.213]\}$.

Then, following the above method with splitting and merging intervals, we have

1. Starting from I_1,
   (a) Step 1: I_1 mapps to $I_3 = \{w_y \leq 0.416, v - w_x \in [-0.162, 0.068], vw_x \in [0.55, 1.12131]\}$.
   (b) Step 2: Splitting I_3, we have
      i. $I_4 = \{w_y \leq 0.416, v - w_x \in [-0.162, 0.068], vw_x \in [0.55, 0.8]\}$.
      ii. $I_5 = \{w_y \leq 0.416, v - w_x \in [-0.162, 0.068], vw_x \in [0.8, 0.9]\}$.
      iii. $I_6 = \{w_y \leq 0.416, v - w_x \in [-0.162, 0.068], vw_x \in [0.9, 1.0]\}$.
      iv. $I_7 = \{w_y \leq 0.416, v - w_x \in [-0.162, 0.068], vw_x \in [1.0, 1.12131]\}$.
      Then, we have
      i. I_4 mapps to
         $I_8 = \{w_y \leq 0.214, v - w_x \in [-0.309, 0.0545], vw_x \in [0.942, 1.25786]\}$.
      ii. I_5 mapps to
         $I_9 = \{w_y \leq 0.0966, v - w_x \in [-0.335, 0.0613], vw_x \in [0.880, 1.19649]\}$.
      iii. I_6 mapps to
         $I_{10} = \{w_y \leq 0.0756, v - w_x \in [-0.362, 0.068], vw_x \in [0.777894, 1.11178]\}$.
      iv. I_7 mapps to
         $I_{11} = \{w_y \leq 0.134, v - w_x \in [-0.394, 0.0782], vw_x \in [0.595, 1]\}$.
   (c) Step 3: Splitting and merging I_8, I_9, I_{10}, I_{11}, we have
      i. $I_{12} = \{w_y \leq 0.134, v - w_x \in [-0.394, 0.078], vw_x \in [0.595, 0.777]\}$.
      ii. $I_{13} = \{w_y \leq 0.214, v - w_x \in [-0.394, 0.078], vw_x \in [0.777, 1]\}$.
      iii. $I_{14} = \{w_y \leq 0.214, v - w_x \in [-0.362, 0.068], vw_x \in [1, 1.11178]\}$.
      iv. $I_{15} = \{w_y \leq 0.214, v - w_x \in [-0.309, 0.061], vw_x \in [1.11178, 1.25786]\}$.
      Then, we have
      i. I_{12} mapps to
         $I_{16} = \{w_y \leq 0.0372, v - w_x \in [-0.317, 0.061], vw_x \in [1.14493, 1.31246]\}$.
      ii. I_{13} mapps to
         $I_{17} = \{w_y \leq 0.0432, v - w_x \in [-0.448, 0.078], vw_x \in [0.943633, 1.24393]\}$.
      iii. I_{14} mapps to
         $I_{18} = \{w_y \leq 0.0662, v - w_x \in [-0.462, 0.077], vw_x \in [0.77846, 1]\}$.
      iv. I_{15} mapps to
         $I_{20} = \{w_y \leq 0.0998, v - w_x \in [-0.456, 0.0785], vw_x \in [0.550, 0.878]\}$.

2. Starting from I_2,
   (a) Step 1: I_2 mapps to $I_{21} = \{w_y \leq 0.332, v - w_x \in [-0.205, 0.114], vw_x \in [0.864, 1.25894]\}$
   (b) Step 2: Splitting I_{21}, we have
      i. $I_{22} = \{w_y \leq 0.332, v - w_x \in [-0.205, 0.114], vw_x \in [0.864, 1]\}$.
      ii. $I_{23} = \{w_y \leq 0.332, v - w_x \in [-0.205, 0.114], vw_x \in [1, 1.125894]\}$.

      Then, we have
      i. I_{22} mapps to
         $I_{24} = \{w_y \leq 0.081, v - w_x \in [-0.336, 0.114], vw_x \in [0.858, 1.14813]\}$.
      ii. I_{23} mapps to
         $I_{25} = \{w_y \leq 0.184, v - w_x \in [-0.409, 0.148], vw_x \in [0.463, 1]\}$.
   (c) Step 3: Splitting and merging I_{24}, I_{25}, we have
      i. $I_{26} = \{w_y \leq 0.184, v - w_x \in [-0.409, 0.148], vw_x \in [0.463, 1]\}$.
      ii. $I_{27} = \{w_y \leq 0.081, v - w_x \in [-0.336, 0.114], vw_x \in [1, 1.14813]\}$.

      Then, we have
      i. I_{26} mapps to
         $I_{28} = \{w_y \leq 0.083, v - w_x \in [-0.452, 0.148], vw_x \in [0.952783, 1.31778]\}$.
      ii. I_{27} mapps to
         $I_{29} = \{w_y \leq 0.034, v - w_x \in [-0.399, 0.133], vw_x \in [0.777, 1]\}$.

3. Starting from I_{30},
   (a) Step 1: I_{30} mapps to $I_{31} = \{w_y \leq 0.44, v - w_x \in [-0.124, 0.037], vw_x \in [0.422, 0.767]\}$
   (b) Step 2: Splitting I_{31}, we have
      i. $I_{32} = \{w_y \leq 0.44, v - w_x \in [-0.124, 0.037], vw_x \in [0.422, 0.5]\}$.
      ii. $I_{33} = \{w_y \leq 0.44, v - w_x \in [-0.124, 0.037], vw_x \in [0.5, 0.6]\}$.
      iii. $I_{34} = \{w_y \leq 0.44, v - w_x \in [-0.124, 0.037], vw_x \in [0.6, 0.767]\}$.

      Then, we have
      i. I_{32} mapps to
         $I_{35} = \{w_y \leq 0.301, v - w_x \in [-0.218, 0.0185], vw_x \in [0.901, 1.20971]\}$.
      ii. I_{33} mapps to
         $I_{36} = \{w_y \leq 0.262, v - w_x \in [-0.245, 0.023], vw_x \in [0.96322, 1.25093]\}$.
      iii. I_{34} mapps to
         $I_{37} = \{w_y \leq 0.213, v - w_x \in [-0.288, 0.029], vw_x \in [0.947, 1.25345]\}$.
   (c) Step 3: Splitting and merging I_{35}, I_{36}, I_{37}, we have
      i. $I_{38} = \{w_y \leq 0.301, v - w_x \in [-0.288, 0.029], vw_x \in [0.901, 1]\}$.
      ii. $I_{39} = \{w_y \leq 0.301, v - w_x \in [-0.288, 0.029], vw_x \in [1, 1.1]\}$.
      iii. $I_{40} = \{w_y \leq 0.301, v - w_x \in [-0.288, 0.029], vw_x \in [1.1, 1.25093]\}$.
      iv. $I_{41} = \{w_y \leq 0.262, v - w_x \in [-0.245, 0.029], vw_x \in [1.25093, 1.25345]\}$.

      Then, we have
      i. I_{38} mapps to
         $I_{42} = \{w_y \leq 0.0404, v - w_x \in [-0.392, 0.029], vw_x \in [0.888, 1.11696]\}$.
      ii. I_{39} mapps to
         $I_{43} = \{w_y \leq 0.0740, v - w_x \in [-0.428, 0.033], vw_x \in [0.741, 1]\}$.
      iii. I_{40} mapps to
         $I_{44} = \{w_y \leq 0.125, v - w_x \in [-0.482, 0.038], vw_x \in [0.497, 0.891]\}$.
      iv. I_{41} mapps to
         $I_{45} = \{w_y \leq 0.109, v - w_x \in [-0.400, 0.038], vw_x \in [0.534, 0.702]\}$.
   (d) Step 4: Splitting and merging $I_{42}, I_{43}, I_{44}, I_{45}$, we have
      i. $I_{46} = \{w_y \leq 0.125, v - w_x \in [-0.482, 0.038], vw_x \in [0.497, 0.891]\}$.
      ii. $I_{47} = \{w_y \leq 0.074, v - w_x \in [-0.428, 0.033], vw_x \in [0.891, 1]\}$.
      iii. $I_{48} = \{w_y \leq 0.041, v - w_x \in [-0.40, 0.029], vw_x \in [1, 1.11696]\}$.

      Then, we have

i. I_{46} mapps to
$I_{49} = \{w_y \leq 0.0424, v - w_x \in [-0.442, 0.034], vw_x \in [1.07853, 1.34708]\}$.
ii. I_{47} mapps to
$I_{50} = \{w_y \leq 0.0110, v - w_x \in [-0.435, 0.033], vw_x \in [0.993, 1.13943]\}$.
iii. I_{48} mapps to
$I_{51} = \{w_y \leq 0.0109, v - w_x \in [-0.454, 0.033], vw_x \in [0.497, 0.891]\}$.

For Case (II):

Now we are to compute the ranges starting from the interval where $I = \{w_y \leq 0.37, v - w_x \in [-0.017, 0.17], vw_x \in [\frac{1}{4\pi}, \frac{1}{2\pi}]\}$. First, we denote it as

1. $I_{52} = \{w_y \leq 0.37, v - w_x \in [-0.017, 0.17], vw_x \in [\frac{1}{4\pi}, \frac{1}{2\pi}]$.

Then, following the above method with splitting and merging intervals, we have

1. Starting from I_{52},
   (a) Step 1: I_{52} mapps to $I_{53} = \{w_y \leq 0.37, v - w_x \in [-0.079, 0.0271], vw_x \in [0.222, 0.616]\}$.
   (b) Step 2: I_{53} mapps to $I_{54} = \{w_y \leq 0.343, v - w_x \in [-0.171, 0.017], vw_x \in [0.621, 1.24894]\}$.
   (c) Step 3: Splitting I_{54}, we have
      i. $I_{55} = \{w_y \leq 0.343, v - w_x \in [-0.171, 0.017], vw_x \in [0.621, 1\}$.
      ii. $I_{56} = \{w_y \leq 0.343, v - w_x \in [-0.171, 0.017], vw_x \in [1, 1.24894]\}$.
      Then, we have
      i. I_{55} mapps to
      $I_{57} = \{w_y \leq 0.150, v - w_x \in [-0.305, 0.017], vw_x \in [0.840, 1.25908]\}$.
      ii. I_{56} mapps to
      $I_{58} = \{w_y \leq 0.137, v - w_x \in [-0.367, 0.022], vw_x \in [0.472, 1]\}$.
   (d) Step 4: Splitting and merging I_{57}, I_{58}, we have
      i.
      ii. $I_{59} = \{w_y \leq 0.150, v - w_x \in [-0.367, 0.022], vw_x \in [0.472, 1\}$.
      iii.
      iv. $I_{60} = \{w_y \leq 0.150, v - w_x \in [-0.305, 0.017], vw_x \in [1, 1.25908\}$.
      Then, we have
      i. I_{59} mapps to
      $I_{61} = \{w_y \leq 0.0705, v - w_x \in [-0.393, 0.022], vw_x \in [0.971, 1.304]\}$.
      ii. I_{60} mapps to
      $I_{62} = \{w_y \leq 0.0613, v - w_x \in [-0.421, 0.0219], vw_x \in [0.583, 1]\}$.

For both Cases (I, II):

From $I_{16-20}, I_{28}, I_{29}, I_{49-51}, I_{61}, I_{62}$, we can see that it has fallen into an interval $I_f = \{w_y < 0.1, v - w_x \in [-0.462, 0.148], vw_x \in [0.497, 1.34078]\}$. Something special here is that w_y has been much smaller than w_x. More broadly, let's define an interval I_s generated by $I_g = \{w_y = 0, v - w_x \in [-0.464, 0.148], vw_x \in [1, 1.5]\}$. Here "generated" means

$$I_s = \bigcup_{T \geq t} \{(v^{(T)}, w_x^{(T)}, w_y^{(T)}) | (v_t^{(t)}, w_x^{(t)}, w_y^{(t)}) \in I_g\}. \tag{99}$$

Then each element $(v, w_x, w_y) \in I_s$ has the following properties:

1. $w_y = 0$.
2. $vw_x \in [0.181, 1.5]$.
3. If $vw_x \leq 1$, then $v - w_x \in [-0.735, 0.23]$. If $vw_x > 1$, then $v - w_x \in [-0.474, 0.148]$.

The first property is obvious. The third can be proven as follows: for each element $(v, w_x, w_y) \in I_g$, it has $v^{(t+1)} - w_x^{(t+1)} = (v - w_x)\,(1 + K(vw_x - 1))$, where the ratio $1 + K(vw_x - 1) \in [1, 1 + 1.1(1.5 - 1)]$ when $vw_x \in [1, 1.5]$. Furthermore, in the proven 2-D case, we have shown that "if $vw_x > 1$ with some mild conditions, then $\frac{v^{(t+2)} - w_x^{(t+2)}}{v - w_x} \in (-1, 1)$". Actually it can be tighter as $\frac{v^{(t+2)} - w_x^{(t+2)}}{v - w_x} \in (-0.2, 1)$ because here $K \leq 1.1$ while the original bound is for $K \leq 1.5$. The condition of bounded $|v - w_x|$ can also be verified, the purpose of which is to keep v, w_x always positive. Then the bound $[-0.2, 1]$ will tell $v - w_x \in [-0.474, 0.148]$ on $vw_x \geq 1$, because

$$\frac{0.148}{0.474} > 0.2, \qquad \frac{0.474}{0.148} > 0.2.$$

For the second property, the left bound can be verified as

$$\min_{x \in [1, 1.5]} x(1 + 1.1(1 - x))^2 + 1.1(1 - x) \cdot 0.474^2 = \left(x(1 + 1.1(1 - x))^2 + 1.1(1 - x) \cdot 0.474^2 \right) \Big|_{x=1.5} \geq 0.181.$$

The right bound can be verified as

$$\max_{x \in [0, 1]} x(1 + 1.1(1 - x))^2 + 1.1(1 - x) * 0.735^2 < 1.5.$$

After proving these three properties, we would like to bound how far I_f is away from I_s. More precisely, the distance is measured by w_y. We are going to show w_y decays exponentially.

Remind the update rules in (66, 67). Denote $\gamma = \frac{1}{\pi}(\arctan(\alpha) - \frac{w_x w_y}{\|w\|^2})$ again and $\delta = K \frac{w_y^2}{v}(-v^2 + \frac{v w_y}{\pi \|w\|^2})$, then it is

$$\Delta v = K w_x(-v w_x + 1) - K w_x \gamma + \delta, \tag{100}$$

$$\Delta w_x = K v(-v w_x + 1) - K v \gamma, \tag{101}$$

$$\delta \in [-K v w_y^2, 0]. \tag{102}$$

Note that both γ and δ are very small, so we are to show their effects separately, which is enough to be a good approximation.

Consider an iteration where $v^{(t)} w_x^{(t)} > 1$ and the corresponding $\gamma^{(t)}$. Let's denote $v^{(t+1)}, w_x^{(t+1)}$ as the next parameters with **no corruption** from $\gamma^{(t)}$. Similarly, we denote $\hat{v}^{(t+1)}, \hat{w_x}^{(t+1)}$ are corrupted with $\gamma^{(t)}$. From the 2-D analysis, we know

$$\frac{v^{(t+2)} - w_x^{(t+2)}}{v^{(t)} - w_x^{(t)}} = (1 + K(v^{(t)} w_x^{(t)} - 1))(1 + K(v^{(t+1)} w_x^{(t+1)} - 1)) < 1. \tag{103}$$

We would like to show, with a small $\gamma^{(t)}$ and ignoring δ,

$$\frac{\hat{v}^{(t+2)} - \hat{w_x}^{(t+2)}}{v^{(t)} - w_x^{(t)}} = (1 + K(v^{(t)} w_x^{(t)} - 1 + \gamma^{(t)}))(1 + K(\hat{v}^{(t+1)} \hat{w_x}^{(t+1)} - 1 + \gamma^{(t+1)})) \lessapprox 1, \tag{104}$$

where $\gamma^{(t+1)}$ is in time $(t + 1)$ accordingly. The difference of LHS of the above two expressions turns out to be

$$\begin{aligned}
(104) - (103) &= K\gamma^{(t)}(1 + K(v^{(t+1)} w_x^{(t+1)} - 1)) \\
&\quad + (1 + K(v^{(t)} w_x^{(t)} - 1)) K(\hat{v}^{(t+1)} \hat{w_x}^{(t+1)} - v^{(t+1)} w_x^{(t+1)} + \gamma^{(t+1)}) + \mathcal{O}(\gamma^2) \\
&= K\gamma^{(t)}(1 + K(v^{(t+1)} w_x^{(t+1)} - 1)) \\
&\quad + K(1 + K(v^{(t)} w_x^{(t)} - 1))(-K(v^{(t)})^2 \gamma^{(t)} - K(w_x^{(t)})^2 \gamma^{(t)} + \gamma^{(t+1)}) + \mathcal{O}(\gamma^2) \\
&\leq K\gamma^{(t)} \left(1 + (1 + K(v^{(t)} w_x^{(t)} - 1)) \big(-K(v^{(t)})^2 - K(w_x^{(t)})^2 + \frac{\gamma^{(t+1)}}{\gamma^{(t)}} \big) \right) + \mathcal{O}(\gamma^2) \\
&\leq K\gamma^{(t)} \left(1 + (1 + K(v^{(t)} w_x^{(t)} - 1)) \big(-2K v^{(t)} w_x^{(t)} + \frac{\gamma^{(t+1)}}{\gamma^{(t)}} \big) \right) + \mathcal{O}(\gamma^2).
\end{aligned} \tag{105}$$

Since $\frac{\Delta w_x}{w_x} = K\frac{v}{w_x}(-vw_x + 1 - \gamma)$, we have

$$\frac{w_x^{(t+1)}}{w_x^{(t)}} = 1 + K\frac{v^{(t)}}{w_x^{(t)}}(-v^{(t)}w_x^{(t)} + 1 - \gamma^{(t)}) < 1. \tag{106}$$

Also we have

$$\frac{\gamma^{(t+1)}}{\gamma^{(t)}} = \frac{\arctan(\frac{w_y^{(t+1)}}{w_x^{(t+1)}}) - \frac{w_x^{(t+1)}w_y^{(t+1)}}{\left\|w^{(t+1)}\right\|^2}}{\arctan(\frac{w_y^{(t)}}{w_x^{(t)}}) - \frac{w_x^{(t)}w_y^{(t)}}{\left\|w^{(t)}\right\|^2}}. \tag{107}$$

Since $w_y^{(t+1)} \leq w_y^{(t)}$ and

$$\frac{\arctan(mx) - \frac{mx}{1+m^2x^2}}{\arctan(x) - \frac{x}{1+x^2}} \leq m^3, \text{ for any } m > 0, x > 0, \tag{108}$$

we have

$$\frac{\gamma^{(t+1)}}{\gamma^{(t)}} \leq \frac{1}{\left(1 + K\frac{v^{(t)}}{w_x^{(t)}}(-v^{(t)}w_x^{(t)} + 1 - \gamma^{(t)})\right)^3}. \tag{109}$$

For general $vw_x \in (1, 1.5]$, (109) holds as

$$\frac{\gamma^{(t+1)}}{\gamma^{(t)}} \lessapprox \frac{1}{(1 + 1.1\frac{\sqrt{1+0.074^2}+0.074}{\sqrt{1+0.074^2}-0.074}(-1.5+1))^3} \leq 22. \tag{110}$$

Since $1 + K(v^{(t)}w_x^{(t)} - 1) \leq 1 + 1.1 * 0.5 = 1.55$, it is fair to say

$$(104) - (103) \lessapprox K\gamma^{(t)}(1 + 1.55 * (-2 + 22)) + \mathcal{O}(\gamma^2) = 35.2\gamma^{(t)} + \mathcal{O}(\gamma^2). \tag{111}$$

Actually $\gamma^{(t)}$ is bounded by

$$\frac{w_y^{(t)}}{w_x^{(t)}} \leq \frac{0.099}{\sqrt{1 + 0.074^2} - 0.074} = 0.1066, \tag{112}$$

$$\gamma^{(t)} \leq \frac{\arctan(x) - \frac{x}{1+x^2}}{\pi} \leq 2.6 \times 10^{-4}. \tag{113}$$

As a result,

$$(104) - (103) \lessapprox 0.0084. \tag{114}$$

Note that this small value is very easy to cover in (103), requiring

$$1 - \frac{v^{(t+2)} - w_x^{(t+2)}}{v^{(t)} - w_x^{(t)}} \geq 0.0084, \tag{115}$$

except when vw_x is pretty close to 1. When $vw_x \to 1$, from the analysis of 2-D case, (derived from the case of $x_{t+1}y_{t+1} \geq x_s^2$)

$$1 - \frac{v^{(t+2)} - w_x^{(t+2)}}{v^{(t)} - w_x^{(t)}} \geq (2K - 2)(v^{(t)}w_x^{(t)} - 1). \tag{116}$$

For $(104) - (103)$, denote a function $p(x)$ as

$$p(x) = 1 + (1 + Kx)\left(-2K(x+1) + \frac{1}{\left(1 + K\frac{v}{w_x}(-x)\right)^3}\right), \tag{117}$$

where $x = v^{(t)}w_x^{(t)} - 1$ in (105, 109). It is obvious that $p(0) = 1 + (-2K + 1) < 0$. When x is small, it turns out

$$p(x) = -2K + 2 + K\left(-2K - 1 + 3\frac{v^{(t)}}{w_x^{(t)}}\right)x + \mathcal{O}(x^2) \tag{118}$$

As a result, $(104) - (103) < 0$ when $vw_x - 1 = o(K-1)$. What if $vw_x - 1 = \Omega(K-1)$? Actually, we can get a better bound by a more care analysis, as

$$\begin{aligned}\frac{(104)-(103)}{K\gamma^{(t)}} \leq 1 + (1 + K(v^{(t)}w_x^{(t)} - 1))\big(-K(v^{(t)})^2 - K(w_x^{(t)})^2 + \frac{\gamma^{(t+1)}}{\gamma^{(t)}}\big) \\ + K\left[v^{(t)}w_x^{(t)}(1 + K(1 - v^{(t)}w_x^{(t)}))^2 - 1\right],\end{aligned} \tag{119}$$

where the last term is due to $v^{(t+1)}w_x^{(t+1)} \leq v^{(t)}w_x^{(t)}(1+K(1-v^{(t)}w_x^{(t)}))^2$. Hence, with this bound, by expanding the last term, (118) becomes

$$p(x) = -2K + 2 + K\left(-2K - 1 + 3\frac{v^{(t)}}{w_x^{(t)}}\right)x + K(1-2K)x + \mathcal{O}(x^2) \tag{120}$$

$$= -2K + 2 + K\left(-4K + 3\frac{v^{(t)}}{w_x^{(t)}}\right)x + \mathcal{O}(x^2), \tag{121}$$

which is definitely negative because

$$\frac{v^{(t)}}{w_x^{(t)}} \leq \frac{\sqrt{1+0.074^2} + 0.074}{\sqrt{1+0.074^2} - 0.074} < 1.16 < \frac{4}{3}. \tag{122}$$

Meanwhile, we are to prove the δ in (100) will not make $\tilde{I}_s$ make $v - w_x < -0.474$ starting from $v - w_x \geq -0.462$. First, in the region of $\{vw_x \in [1, 1.5], v - w_x \leq 0.148\}$, we have $Kvw_y^2 \leq 1.1 \cdot (\sqrt{1.5 + 0.074^2} + 0.074) * 0.1^2 \leq 0.0144$. Also note that in this region with $v - w_x \geq -0.462$, we have

$$\frac{w_y^{(t+1)}}{w_y} \leq 1 - \frac{\sqrt{1+0.231^2} - 0.231}{\sqrt{1+0.231^2} + 0.231} = 0.37. \tag{123}$$

Hence $Kv(w_y^{(t)})^2 + Kv(w_y^{(t+1)})^2 \leq 0.0144 * (1 + 0.37^2) = 0.0164$. Since $|v^{(t+2)} - w^{(t+2)}| < |v^{(t)} - w^{(t)}|$ if there is no δ, we shall see that there is no need to discuss the case of $v - w_x \geq -0.462 + 0.0164 = -0.4456$ because it still holds $v^{(t+1)} - w_x^{(t+1)} > -0.462$. When $v^{(t)} - w_x^{(t)} \in [-0.462, -0.4456]$, we shall see that in (45), after adding the term of δ in v,

$$\frac{v^{(t+2)} - w_x^{(t+2)}}{v^{(t)} - w_x^{(t)}} \leq 1 - (1 + K(vw_x - 1)) \cdot Kw_x\delta, \tag{124}$$

which means the absolute value of $v - w_x$ decays at least by a margin depending on δ. After multiplying the current difference $v^{(t)} - w_x^{(t)}$ on both side, it gives

$$(v^{(t+2)} - w_x^{(t+2)}) - (v^{(t)} - w_x^{(t)}) \geq v^{(t)}w_x^{(t)}w_x^{(t)}\delta. \tag{125}$$

Note that here $v^{(t+2)} - w_x^{(t+2)}$ does not include $\delta^{(t)}$ and $\delta^{(t+1)}$. As stated above, we have $\frac{\delta^{(t+1)}}{\delta^{(t)}} \leq 0.37^2 \leq 0.16$ due to the decay of w_y. So it is safe to say $\delta^{(t)} + \delta^{(t+1)} \geq 1.16\delta^{(t)}$. Combining with the above inequality, it gives

$$(v^{(t+2)} - w_x^{(t+2)}) - (v^{(t)} - w_x^{(t)}) + \delta^{(t)} + \delta^{(t+1)} \geq (v^{(t)}w_x^{(t)}w_x^{(t)} + 1.16)\delta^{(t)}, \tag{126}$$

where

$$v^{(t)}w_x^{(t)}w_x^{(t)} + 1.16 \leq vw_x \cdot (\sqrt{vw_x + (\frac{0.4456}{2})^2} - \frac{0.4456}{2}) + 1.16 \leq 0.6. \tag{127}$$

Furthermore, from our previous discussion, $w_y^{(t+2)} < w_y^{(t+2)}$ gives that the sum of (126) is bounded by

$$\frac{0.6}{1-0.16}\delta^{(t)} \geq \frac{0.6}{1-0.16} \cdot (-0.0144) \geq -0.0103. \tag{128}$$

Since $-0.474-(-0.462) < -0.0103$, we shall see that the term of δ cannot drive $v - w_x < -0.472$. Note that (126) shall include a factor (< 1) in front of $\delta^{(t)}$, but we have ignored it to show a more aggressive bound.

Therefore, we are able to say an Interval $\hat{I}_s$ generated by I_f also has the following properties: for each element $(v, w_x, w_y) \in \hat{I}_s$,

1. $vw_x \in [0.181, 1.5]$.
2. If $vw_x \leq 1$, then $v - w_x \in [-0.735, 0.23]$. If $vw_x > 1$, then $v - w_x \in [-0.472, 0.148]$.

Then the decreasing ratio of $\Delta w_y / w_y$ is bounded by

$$\frac{\Delta w_y}{w_y} = K\frac{v}{w_x}\left(-vw_x + \frac{1}{\pi}\frac{w_x w_y}{\|w\|^2}\right) \tag{129}$$

$$\in \left[-1.1(\sqrt{1.5+0.074^2}+0.074)^2, -0.030K\right] \tag{130}$$

$$= [-1.87, -0.030K]. \tag{131}$$

Hence, w_y decays with a linear ratio of 0.97 (or $1-0.030K$) at most for Cases (I, II) in stage 2.

For Case (III), in the first step of stage 2, it already has $w_y \leq 0.078$ and $v - w_x \in [-0.017, 0.17]$. So surely it will also converge to I_s.

Here we present the time analysis for Case (III) of both stages. The number of iterations in the first stage is apparently similar to that of case (I, II), as

$$T_1 \leq \log_{2.56}\left\lceil \frac{2.7\psi}{\beta^2} \right\rceil, \tag{132}$$

where $\psi < \frac{1}{4\pi}$ is the value of vw_x in the first step of stage 2. In stage 2, since our target is to find how many steps are necessary to get $vw_x \geq 0.181$, so it is

$$v^{(t+1)}w_x^{(t+1)} \geq v^{(t)}w_x^{(t)}\left(1 - 0.181 + 1 - \frac{\arctan(2-\sqrt{3}) - \frac{2-\sqrt{3}}{1+(2-\sqrt{3})^2}}{\pi} - 1.1w_y^2\right) \tag{133}$$

$$\geq 3.28v^{(t)}w_x^{(t)}. \tag{134}$$

where obviously it still holds $\frac{w_y}{w_x} \leq 2-\sqrt{3}$ and $w_y^2 < 0.1^2$ in stage 2. Since $3.28 > 2.56$, we have the total number of steps to have $vw_x > 0.181$ bounded as

$$\begin{aligned}\left\lceil \log_{2.56}\frac{2.7\psi}{\beta^2}\right\rceil + \left\lceil \log_{3.28}\frac{0.181}{\psi}\right\rceil &\leq \left\lceil \log_{2.56}\frac{0.675}{\pi\beta^2}\right\rceil + \left\lceil \log_{3.28}\frac{0.181}{\frac{1}{4\pi}}\right\rceil + 2 \\ &\leq \left\lceil \log_{2.56}\frac{0.675}{\pi\beta^2}\right\rceil + 3 \\ &< \left\lceil \log_{2.56}\frac{1.35}{\pi\beta^2}\right\rceil + 4,\end{aligned}$$

which is not beyond the bound for Cases (I, II). □

H USEFUL LEMMAS

Lemma 7. *Assume $a \cdot \Delta a \geq b \cdot \Delta b$ and $a \geq b$. All of $a, b, \Delta a, \Delta b$ are positive. If $\Delta b \leq a$, then $a + \Delta a \geq b + \Delta b$.*

Proof. $(a+\Delta a) - (b+\Delta b) \geq a + b\frac{\Delta b}{a} - b - \Delta b = (\frac{\Delta b}{a} - 1)(b-a) \geq 0.$ □

I ADDITIONAL EXPERIMENTS

We perform two experiments in relatively higher dimension settings. We are to show two observations that coincides with our discussions in the low dimension:

Observation 1: GD beyond EoS derives to flatter minima.

Observation 2: GD beyond EoS is in a similar style with the low dimension.

I.1 2-LAYER HIGH-DIM HOMOGENEOUS RELU NNS WITH PLANTED TEACHER NEURONS

We conduct a synthetic experiment in the high-dimension teacher-student framework. The teacher network is in the form of

$$y|x := f_{\text{teacher}}(x;\tilde{\theta}) = \sum_{i=1}^{16} \mathsf{ReLU}(\mathbf{e}_i^\top x), \tag{135}$$

where $x \in \mathbb{R}^{16}$ and $\mathbf{e_i}$ is the i-th vector in the standard basis of $\mathbb{R}^{16}$. The student and the loss are in forms of

$$f(x;\theta) = \sum_{i=1}^{16} v_i \cdot \mathsf{ReLU}(w_i^\top x), \tag{136}$$

$$L(\theta;\tilde{\theta}) = \frac{1}{m}\sum_{i}^{16} \left(f(x;\theta) - y|x_i\right)^2. \tag{137}$$

Apparently, the global minimum manifold contains the following set $\mathcal{M}$ as (w.l.o.g., ignoring any permutation)

$$\mathcal{M} = \{(v_i, w_i)_{i=1}^{16} \mid \forall i \in [16], w_i = k_i \cdot \mathbf{e}_i, v_i = \frac{1}{k_i}, k_i > 0\}. \tag{138}$$

However, different choices of $\{k_i\}_{i=1}^{16}$ induce different extents of sharpness around each minima. Our aim is to show that **GD with a large learning rate beyond the edge of stability drives to the flattest minima from sharper minima**.

Initialization. We initialize all student neurons directionally aligned with the teachers as $w_i \parallel \mathbf{e}_i$ but choose various k_i, as $k_i = 1 + 0.0625(i-1)$. Obviously, such a choice of $\{k_i\}_{i=1}^{16}$ is not at the flattest minima, due to the isotropy of teacher neurons. Also we add small noise to w_i to make the training start closely (but not exactly) from a sharp minima, as

$$w_i = k_i \cdot (\mathbf{e}_i + 0.01\epsilon), \quad \epsilon \sim \mathcal{N}(0, I). \tag{139}$$

Data. We uniformly sample 10000 data points from the unit sphere $\mathcal{S}^{15}$.

Training. We run gradient descent with two learning rates $\eta_1 = 0.5, \eta_2 = 2.6$. Later we will show with experiments that the EoS threshold of learning rate is around 2.5, so η_2 is beyond the edge of stability. GD with these two learning rates starts from the same initialization for 100 epochs. Then we extend another 20 epochs with learning rate decay to 0.5 from 2.6 for the learning-rate case.

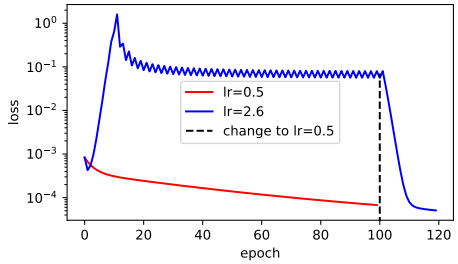
Results. All results are provided in Figure 4. Both Figure 4 (a, b) present the gap between these two trajectories, where GD with a small learning rate stays around the sharp minima, while that with a larger one drives to flatter minima. Then GD stably oscillates around the flatter minima.

Meanwhile, from Figure 4 (b), when we decrease the learning rate from 2.6 to 0.5 after 100 epochs, GD converges to a nearby minima which is significantly flatter, compared with that of lr=0.5.

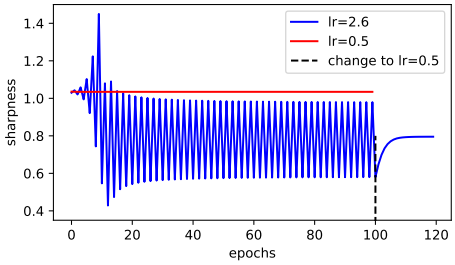
Figure 4 (c) provides a more detailed view of $\frac{\|w_i\|}{v_i}$ for all 16 neurons. All neurons with lr=0.5 stay at the original ratio k_i^2. But those with lr=2.6 all converge to the same ratio around $k^2 = \frac{\|w\|}{v} = 1.21$, as shown in Figure 4 (d). We compute the relationship between the sharpness of global minima in

$\mathcal{M}$ and different choices of k, as shown in Figure 4 (e, f). Actually, $k^2 = 1.21$ is the best choice of $\{k_i\}_{i=1}^{16}$ such that the minima is the flattest.

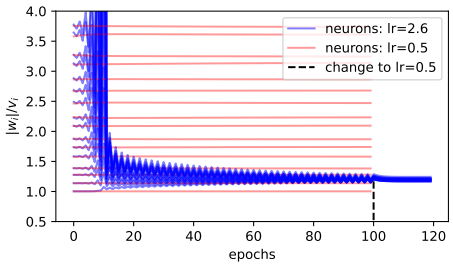
Therefore, we have shown that, **in such a setting of high-dimension teacher-student network, GD beyond the edge of stability drives to the flattest minima**.



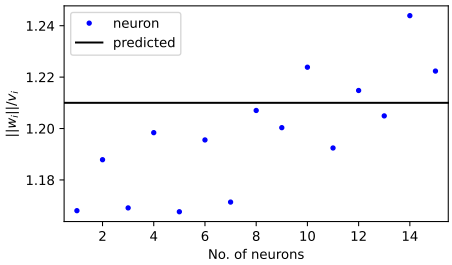
(a) Training loss



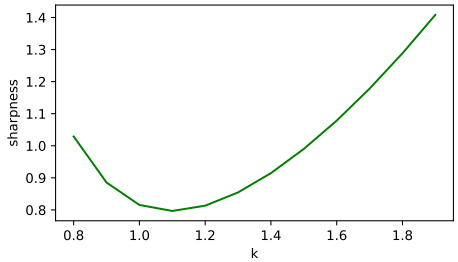
(b) Sharpness



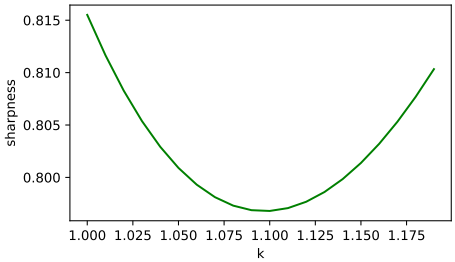
(c) ratio of $\|w_i\|$ and v_i during training



(d) final ratio of $\|w_i\|$ and v_i when lr=2.6



(e) sharpness for different ratio $\sqrt{\frac{\|w_i\|}{v_i}}$



(f) sharpness for different ratio $\sqrt{\frac{\|w_i\|}{v_i}}$ (zoom-in)

Figure 4: Result of 2-layer 16-neuron teacher-student experiment.

I.2 3, 4, 5-LAYER NON-HOMOGENEOUS MLPS ON MNIST

We conduct an experiment on real data to show that **our finding in the low-dimension setting in Theorem 1 is possible to generalize to high-dimensional setting**. More precisely, our goals are to show, when GD is beyond EoS,

1. the oscillation direction (gradient) aligns with the top eigenvector of Hessian.
2. the 1D function at the cross-section of oscillation direction and high-dim loss landscape satisfies the conditions in Theorem 1.

Network, dataset and training. We run 3, 4, 5-layer ReLU MLPs on MNIST LeCun et al. (1998). The networks have 16 neurons in each layer. To make it easier to compute high-order derivatives, we simplify the dataset by 1) only using 2000 images from class 0 and 1, and 2) only using significant input channels where the standard deviation over the dataset is at least 110, which makes the network input dimension as 79. We train the networks using MSE loss subjected to GD with large learning rates $\eta = 0.5, 0.4, 0.35$ and a small rate $\eta = 0.1$ (for 3-layer). Note that the larger ones are beyond EoS.

Definition 2 (line search minima)**.** *Consider a function f, learning rate η and a point $x \in domain(f)$. We call $\tilde{x}$ as the line search minima of x if*

$$\tilde{x} = x - c^* \cdot \eta \nabla f(x), \tag{140}$$

$$c^* = argmin_{c \in [0,1]} f\left(x - c \cdot \eta \nabla f(x)\right). \tag{141}$$

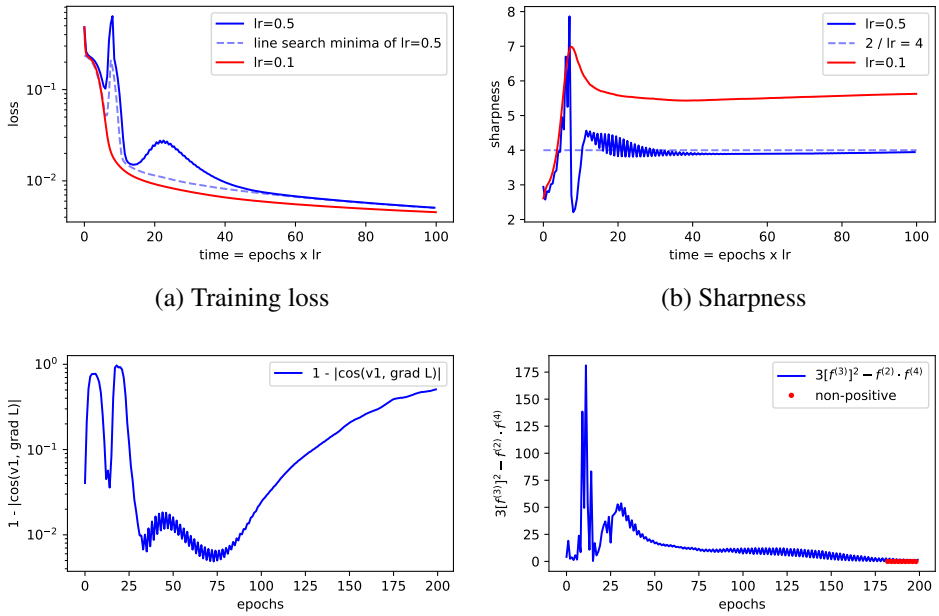
The line search minima $\tilde{x}$ can interpreted as the lowest point on the 1D function induced by the gradient at x. If GD is beyond EoS, $\tilde{x}$ stays in the valley below the oscillation of x.

Results. All results are presented in Figure 5, 6 and 7.

Take the 3-layer as an example. From Figure 5 (a, b), GD is beyond EoS during epochs 10-14 and 21-60. For these epochs, cosine similarity between the top Hessian eigenvector v_1 and the gradient is pretty close to 1, as shown in Figure 5 (c), which verifies our goal 1.

In Figure 5 (d), we compute $3[f^{(3)}]^2 - f^{(2)} f^{(4)}$ at line search minima along training, which is required to be positive in Theorem 1 to allow stable oscillation. Then it turns out most points have $3[f^{(3)}]^2 - f^{(2)} f^{(4)} > 0$ except a few points, all of which are not in the EoS regime, and these few exceptional points might be due to approximation error to compute the fourth-order derivative since their negativity is quite small. This verifies our goal 2.

Both the above arguments are the same in the cases of 4 and 5 layers as shown in Figure 6 and 7.



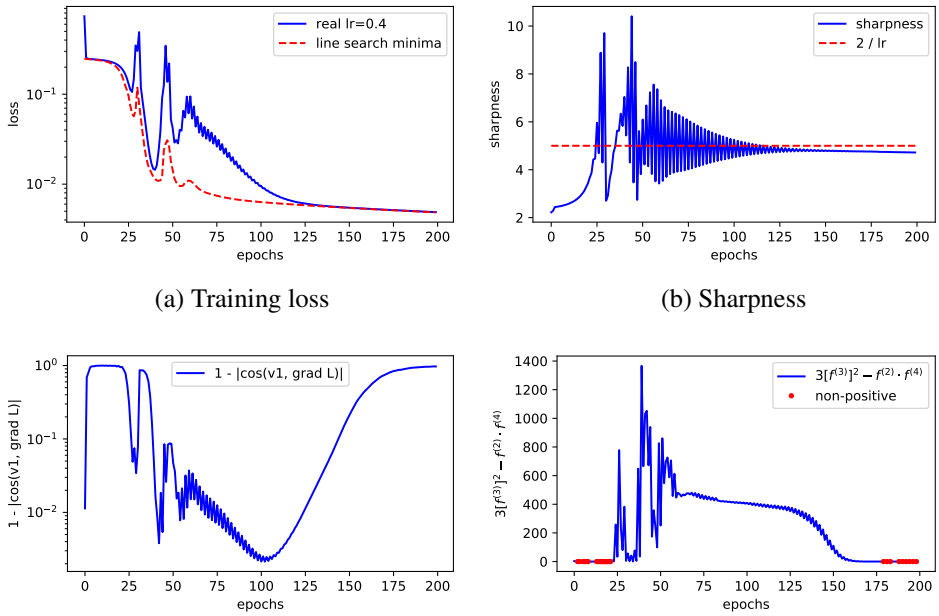
(a) Training loss

(b) Sharpness

(c) similarity of gradient and top eig-vector v_1

(d) $3[f^{(3)}]^2 - f^{(2)} f^{(4)}$ at line search minima

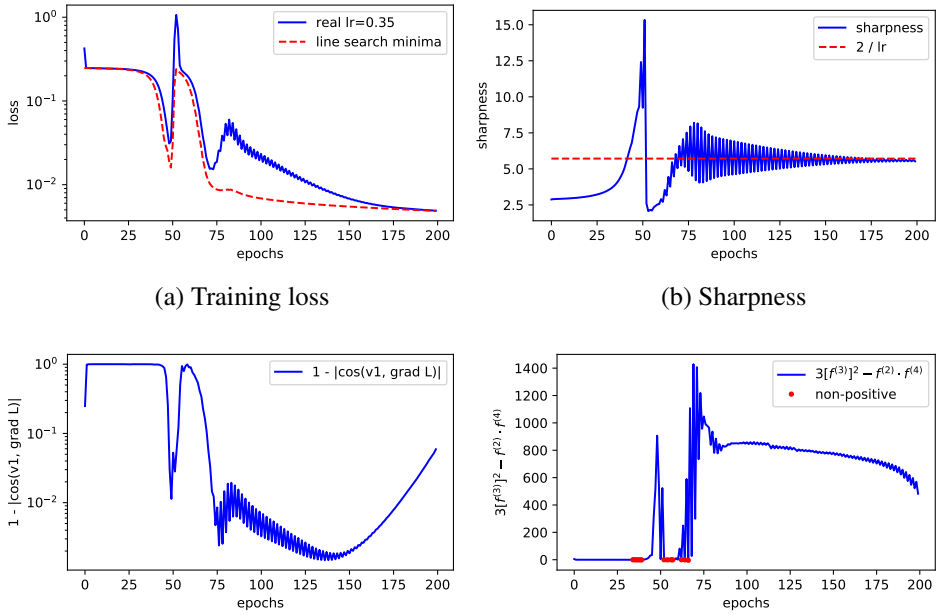
Figure 5: Result of **3-layer** ReLU MLPs on MNIST. Both (c) and (d) are for learning rate as 0.5.



(a) Training loss (b) Sharpness

(c) similarity of gradient and top eig-vector v_1 (d) $3[f^{(3)}]^2 - f^{(2)}f^{(4)}$ at line search minima

Figure 6: Result of **4-layer** ReLU MLPs on MNIST.



(a) Training loss (b) Sharpness

(c) similarity of gradient and top eig-vector v_1 (d) $3[f^{(3)}]^2 - f^{(2)}f^{(4)}$ at line search minima

Figure 7: Result of **5-layer** ReLU MLPs on MNIST.

J ILLUSTRATION OF PERIOD-2 AND PERIOD-4 ORBITS

In the proof of local convergence to the period-2 orbit in (28), we give a bound of learning rate as $\sqrt{5}-1 \approx 1.236$. Local convergence is guaranteed if the learning rate is smaller than it. Conversely, if the learning rate is larger than it, although the period-2 orbit still exists, GD starting from a point infinitesimally close to the orbit still escapes from it. This is when GD converges to a higher-order orbit.

Figure 8 precisely shows the effectiveness of such a bound where GD converges to the period-2 orbit when $\eta = 1.235 < \sqrt{5}-1$ and a period-4 orbit when $\eta = 1.237 > \sqrt{5}-1$.

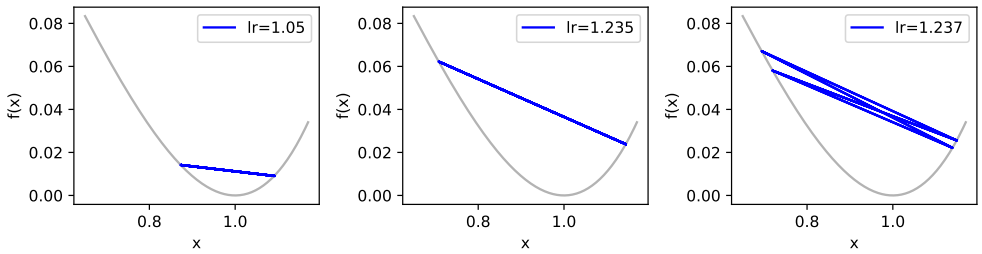


Figure 8: The convergent orbits of GD on $f(x) = \frac{1}{4}(x^2-1)^2$ with learning rate=1.05, 1.235 and 1.237. The first two smaller learning rates drive to period-2 orbits while the last one goes to an period-4 orbit. The significant bound between period-2 and period-4 is predicted by our proof in (28), as $\sqrt{5}-1 \approx 1.236$.