Good regularity creates large learning rate implicit biases: edge of stability, balancing, and catapult

Yuqing Wang **Zhenghao Xu** **Tuo Zhao** **Molei Tao**
{ywang3398, zhenghaoxu, tourzhao, mtao}@gatech.edu
Georgia Institute of Technology
Atlanta, GA 30332

Abstract

Large learning rates, when applied to gradient descent for nonconvex optimization, yield various implicit biases including edge of stability [1], balancing [2], and catapult [3]. These phenomena cannot be well explained by classical optimization theory. Significant theoretical progress has been made to understand these implicit biases, but it remains unclear for which objective functions would they occur. This paper provides an initial step in answering this question, showing that these implicit biases are different tips of the same iceberg. To establish these results, we develop a global convergence theory under large learning rates for two examples of nonconvex functions without global smoothness, departing from typical assumptions in traditional analyses. Specifically, these phenomena are more likely to occur when the optimization objective function has good regularity. This regularity, together with gradient descent using a large learning rate that favors flatter regions, result in these nontrivial dynamical behaviors. We also discuss the implications on training neural networks, where different losses and activations can affect regularity and lead to highly varied training dynamics.

1 Introduction

Large learning rates are often employed in deep learning practices, which is believed to improve training efficiency and generalization [4, 3, 5, 6], while they are still theoretically under-explored. A prevalent hypothesis states that using large learning rates results in the emergence of "flat minima," which in turn leads to better generalization. This belief has served as a catalyst for many novel insights examining the 'sharpness' of the solution under large learning rates, i.e., the largest eigenvalue of the Hessian matrix associated with the objective function. This perspective has given rise to diverse implicit biases, including edge of stability (EoS) [1], balancing [2], and catapult [3]. Collectively, these phenomena will hereafter be referred to as as *large learning rate phenomena*.

There are a lot of theoretical works that analyze EoS and balancing for specific objective functions [7–16]. However, it was unclear whether or how these phenomena are related to each other, or what trigger them. Two important but unsolved questions are the following:

When and why do EoS and other large learning rate phenomena occur?

To answer these questions, we consider $\min_u f(u)$ for two example functions [1] optimized by gradient descent (GD) $u_{k+1} = u_k - h\nabla f(u_k)$, where $h > 0$ is the learning rate. We analyze the convergence of GD under large learning rate, where $\frac{2}{L} < h \lesssim \frac{4}{L}$ and L is the local Lipschitz constant of ∇f (see Sec. 2.3 and 2.4 for detailed definitions and discussions) and obtain the following theoretical results:

• **About 'when'.** Under the two example functions, we demonstrate that large learning rate phenomena depend on the regularity of the objective functions. Roughly speaking, in the case of functions

[1] A longer version of this paper that studies more general functions can be found in [17].

with good regularity, EoS and balancing phenomena are more prone to appear. However, when dealing with functions of poor regularity, these two phenomena are more likely to disappear.

- **About 'why'.** Our theoretical analysis reveals that a crucial characteristics of these phenomena is the ability of large learning rates to steer GD towards flatter regions. In other words, the sharpness along GD iterations is influenced and controlled by how large the learning rate is. In fact, upon the convergence of GD, which is nontrivial due to large learning rate but proved as a groundwork for our theory, the limiting sharpness is within $\tilde{\mathcal{O}}(h)$ distance below $2/h$ for the function of good regularity. For the bad one, it is bounded by $1/h$ and we name this nontrivial phenomenon as *one-sided stability*. Moreover, when the objective function has good regularity, GD is guaranteed to enter a region with sharpness less than its limiting sharpness (which is an early stage of progressive sharpening), and then 'crawl' up in sharpness till $\approx 2/h$ (near the edge of stability).

Notation. We use $\lesssim$ such that $a \lesssim b$ means $a < b + \epsilon$ for some small $\epsilon > 0$. We denote S_k to be the largest eigenvalue of Hessian, i.e., sharpness, at kth iteration of GD. We follow the theoretical computer science convention and use $\mathcal{O}(\cdot)$ to indicate the order of a quantity and $\tilde{\mathcal{O}}(\cdot)$ for its order omitting logarithmic dependence. We use $\|\cdot\|_F$ to denote the Frobenius norm of a matrix.

2 Main theory

To better quantify the effect of regularity, we define the following concept:

Definition 1 (Degree of regularity)**.** *Given function $F(s):\mathbb{R}\to\mathbb{R}$, its degree of regularity is*

$$\mathrm{dor}(F) = \inf\left\{n : |F(s)| \le C_1|s|^n, \textit{ for } |s| \ge C_0 \textit{ with some constant } C_0, C_1 > 0\right\}. \tag{1}$$

Inspired by [2, 11, 9], we analyze and compare the following two example functions of x, y

$$f_{\mathrm{good}}(x,y) = F_{\mathrm{good}}(xy) = 2(\log(\exp(xy-1)+1) + \log(\exp(1-xy)+1)), \tag{2}$$

$$f_{\mathrm{bad}}(x,y) = F_{\mathrm{bad}}(xy) = (1-(xy)^3)^2/18. \tag{3}$$

The two functions are designed for a fair comparison with the following properties:

- All the minima of these functions are global minima, located at $xy = 1$;
- The sharpness at any minimizer (x,y) is x^2+y^2.

According to Def. 1, these two functions have different degrees of regularity:

$$\mathrm{dor}(F_{\mathrm{good}}) = 1 \quad < \quad \mathrm{dor}(F_{\mathrm{bad}}) = 6.$$

We prove that for good regularity function $f_{\mathrm{good}}(x,y) = F_{\mathrm{good}}(xy)$ (small dor), EoS and balancing occur, while for bad regularity function $f_{\mathrm{bad}}(x,y) = F_{\mathrm{bad}}(xy)$ (large dor), these phenomena disappear. Before proceeding with the main results, we first describe the two phenomena, EoS and balancing (and also catapult).

2.1 Description of Edge of Stability (EoS) in our framework

EoS is a large learning rate phenomenon. Its original description in [1] contained two stages, namely progressive sharpening, and limiting sharpness stabilization around $2/h$. [9] later observed a third stage before progressive sharpening, which will be referred to as pre-EoS, where the sharpness will first decrease before increasing. A description of the full process is the following:

- **Pre-EoS (de-sharpening).** This stage characterizes the situation where at the very beginning of the iterations, the sharpness decreases sharply before the occurrence of the well-known EoS (see [9]). It does not necessarily come with all the EoS phenomenon and only appears when the initial sharpness is very large. Nevertheless, it helps demonstrate the behavior of GD under large learning rate.

- **Progressive sharpening.** This stage is governed by increasing sharpness. Due to the preparation of the pre-EoS stage, progressive sharpening is guaranteed to start in a relatively flat region (small sharpness), even if GD was initialized in a sharp region. Then, as GD further proceeds, the minimizer that GD eventually converges to will have larger sharpness than the majority of sharpness values along GD trajectory in this stage. Such behavior stems from the good regularity of the functions.

- **Limiting sharpness near $2/h$.** Stability theory of the GD dynamics (see Appendix A) guarantees that the limiting sharpness has to be not exceeding $2/h$, but not necessarily close to $2/h$. It is the good regularity of objective function that will drive the final sharpness towards $2/h$.

2.2 Description of Balancing and Catapult in our framework

The existence of balancing phenomenon is originally proved in matrix factorization problem with objective function $\frac{1}{2}\|A - XY^\top\|_F^2$ [2], for which GD optimization can still converge when LR h exceeds

$2/L$, and its limiting point (X_∞, Y_∞) satisfies $\left| \|X_\infty\|_F - \|Y_\infty\|_F \right|^2 < \frac{2}{c_1 h} - c_2,$ for some $c_1, c_2 > 0$. The larger the learning rate is, the smaller the gap between the magnitudes of X and Y will be, i.e. X and Y will become more balanced when compared to their initials. This is in stark contrast to the behavior observed under small learning rates, where weight discrepancies remain approximately constant [18–20]. Consequently, large learning rates bias GD towards flat minima, even when the optimization is initiated arbitrarily close to a sharp minimum (see Corollary 3.3 in [2] and Sec. 2.3).

There are two phases in the balancing mechanism. The first phase gives rise to *loss catapulting* [3] which is another intriguing empirical observation. In this phase, the loss experiences an initial increase before subsequent decrease, due to GD escaping from the neighborhood of a sharp minimum and then searching for a flatter region that is close to a flat minimum. Additionally, the main mechanism of balancing (the first phase) is similar to **de-sharpening (pre-EoS)** in EoS.

2.3 Good regularity function

In this section, we show that for function with good regularity (2), both EoS and balancing occur when learning rate is large.

Theorem 1. *Assume the initial condition satisfies* $(x_0, y_0) \in \{(x,y) : \sqrt{2} < xy < 4, x^2 + y^2 \gtrsim 48\sqrt[3]{12}\}\backslash\mathcal{B}$*, where* $\mathcal{B}$ *is a Lebesgue measure-0 set. Let the learning rate be* $h = \frac{C}{x_0^2 + y_0^2}$ *for* $2.2 \lesssim C \leq 4$*. Then for* $f_{\text{good}}(x,y)$ (2)*, GD converges to a global minimum* (x_∞, y_∞)*, and:*
- ***EoS*** *(end of pre-EoS; preparation for progressive sharpening): There exists* $N \in \mathbb{N}$*, such that the sharpness at the* N*th iteration satisfies* $S_N \lesssim \frac{1}{4}(6 - C)S_\infty$*, which is* $< S_\infty$ *for all* C *above.*
- ***EoS*** *(Limiting sharpness): The limiting sharpness satisfies* $\frac{2}{h} - \tilde{\mathcal{O}}(h) \leq S_\infty = x_\infty^2 + y_\infty^2 \leq \frac{2}{h}$*.*
- ***Balancing****: The limit* (x_∞, y_∞) *satisfies* $(x_\infty - y_\infty)^2 \leq \frac{2}{C}(x_0 - y_0)^2 + 2(x_0 y_0 - 1)$*.*

Interpretation of EoS results. The above theorem embodies a detailed description of the whole EoS process. More precisely, there is **progressive sharpening**: since there is at least one point along the trajectory with small sharpness (smaller than limiting sharpness), the sharpness will eventually increase to the limiting sharpness. In the end, the **sharpness stabilizes near** $2/h$ within a distance of $\tilde{\mathcal{O}}(h)$. The **pre-EoS (de-sharpening)** can also occur if the initial condition (x_0, y_0) is close to the minima. In this case, the initial sharpness S_0 is approximately $x_0^2 + y_0^2$ and we have

$$S_N \lesssim \frac{1}{4}(6 - C)S_\infty \approx \frac{1}{4}(6 - C)\frac{2}{h} = \left(\frac{3}{C} - \frac{1}{2}\right)(x_0^2 + y_0^2) \approx \left(\frac{3}{C} - \frac{1}{2}\right)S_0 < S_0, \text{ for } 2.2 \lesssim C \leq 4$$

which means the sharpness will first decrease before GD enters the progressive sharpening stage. This also implies that the limiting sharpness is smaller than the initial sharpness, which corresponds to the balancing phenomenon (see the discussion below).

Interpretation of balancing result. The above result states that the limiting difference $(x_\infty - y_\infty)^2$ is upper bounded by its initial. Moreover, if the learning rate h increases (i.e., C increases), the upper bound of $(x_\infty - y_\infty)^2$ will be smaller. For example, when $C = 4$ and $(x_0 - y_0)^2$ are large, we have $(x_\infty - y_\infty)^2 \lesssim \frac{1}{2}(x_0 - y_0)^2$. Indeed, balancing can be used as an explicit characterization of sharpness. To see this, let us first recall that at any minimizer ($xy = 1$), the sharpness is $x^2 + y^2$. Then at that minimizer, we have $(|x| - |y|)^2 = (x - y)^2 = x^2 + y^2 - 2xy = x^2 + y^2 - 2$, i.e., $(x - y)^2$ is equivalent to the sharpness $x^2 + y^2$ up to a constant shift. Therefore, if GD starts near a sharp minimum, large learning rate leads to smaller value of $(x_\infty - y_\infty)^2$, meaning GD converges to a flatter minimum.

Interpretation of large learning rate. Here is how the h's that we considered are large learning rates even when there is no global Lipschitzness: given any h considered in the theorem, we showed GD will converge. Its trajectory is thus in a bounded region. We then consider the maximum of local Lipschitz constant of the gradient over all points in this region, denoted by L. We can show that our h satisfies $h \geq \frac{2}{L}$ and can be approximately $\frac{4}{L}$. Especially, if GD is initialized near the minima, the initial sharpness is $\approx x_0^2 + y_0^2$, which is $\approx L$. Consequently, $h \approx C/L > 2/L$. See more explanation in our longer version.

2.4 Bad regularity function

This section shows that for bad regularity function (3), neither EoS nor balancing occurs even when the learning rate is large.

Theorem 2. *Assume the same initial conditions as Theorem 1. Let the learning rate be* $h = \frac{C}{(x_0^2 + y_0^2 + 4)(x_0 y_0)^4}$ *for* $2 \leq C \leq 3$*. Then for* $f_{\text{bad}}(x,y)$ (3)*, GD converges to a global minimum*

(x_∞, y_∞), *and:*

- ***No EoS****: The limiting sharpness* S_∞ *satisfies* $S_\infty = x_\infty^2 + y_\infty^2 \le \frac{1}{h}$.
- ***No balancing****:* $(x_\infty - y_\infty)^2 \ge (x_0 - y_0)^2 + \min\left\{2(x_0y_0 - 1) - \frac{2C}{3}x_0y_0, -\frac{2C}{12-C}(x_0y_0 - 1)\right\}$.

Interpretation of results. The above theorem shows that the limiting sharpness is $1/h$, which is far below $2/h$, and therefore there is no EoS in this case. We refer to this phenomenon as one-sided stability. Additionally, the limiting difference $(x_\infty - y_\infty)^2$ has a lower bound, implying that the sharpness cannot decrease much at the limit, i.e., GD will not find flatter minimum. Note the lower bound is a monotonically decreasing function w.r.t. C and consequently the learning rate h. This means that larger learning rate can still reduce this lower bound although it cannot reproduce balancing phenomenon.

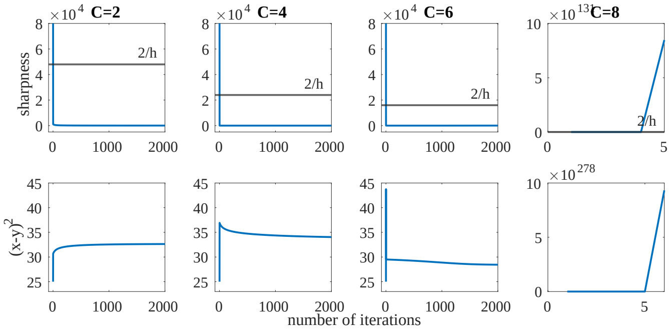
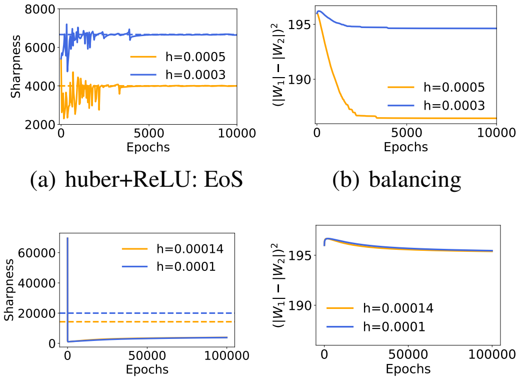


Figure 1: No EoS or balancing for bad regularity no matter what the learning rate is. All the figures share the same initial condition $x_0 = 6, y_0 = 1$.

Larger learning rate cannot help. By Lemma 10, such choice of h is indeed large learning rate, meaning $h > \frac{2}{L}$ and its upper bound is also $\ge \frac{4}{L}$, where L is the local Lipschitz constant of gradient in the bounded region containing the trajectory. Beyond this upper bound, GD may still converge. Nevertheless, the elimination of large learning rate phenomena is independent of learning rate. As is shown in Fig. 1, we implement GD with various learning rates $h = \frac{C}{(x_0^2+y_0^2)(x_0y_0)^4}$, where $C = 2, 4, 6, 8$, until divergence. In all cases, the limiting sharpness is far below $2/h$ and hence there is no EoS; also, we have $(x_\infty - y_\infty)^2 > (x_0 - y_0)^2$ and hence there is no balancing either.



(a) huber+ReLU: EoS (b) balancing

(c) ℓ^2+cubic ReLU: no EoS (d) no balancing

Figure 2: Large learning rate phenomena in neural network models. The dashed lines are $2/h$ for different learning rate h.

3 Neural network implications

To relate the much-more-complicated neural network models to our theory, let's consider a toy example of a 3-layer neural network trained on one data point $(1, 1)$ with linear first layer and fixed last layer (assumed to be 1). Then the training objective function is $f(W_1, W_2) = \mathcal{L}(1, \sigma(W_2W_1))$, where $W_1, W_2^\top \in \mathbb{R}^n$ are the weights and $\mathcal{L}$ is the loss. This function could be rewritten as $f(W_1, W_2) = F(W_1W_2)$ and we have $\mathrm{dor}(F) = \mathrm{dor}(\mathcal{L}(1,\cdot))\mathrm{dor}(\sigma)$. This means the regularity of this objective function depends on two parts: one is the neural network model g whose regularity depends on that of the activation function, and the other is the loss function $\mathcal{L}$. To exemplify such differences, we consider huber loss+ReLU ($\mathrm{dor} = 1$, good regularity) and ℓ^2 loss+cubic ReLU ($\mathrm{dor} = 6$, bad regularity). As is shown in Fig. 2, the former case exhibits large learning rate phenomena (EoS and balancing), while the latter case does not. More details can be found in Appdx.E and our longer version.

References

[1] J. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar, “Gradient descent on neural networks typically occurs at the edge of stability,” in *International Conference on Learning Representations*, 2021.

[2] Y. Wang, M. Chen, T. Zhao, and M. Tao, “Large learning rate tames homogeneity: Convergence and balancing effect,” in *International Conference on Learning Representations*, 2022.

[3] A. Lewkowycz, Y. Bahri, E. Dyer, J. Sohl-Dickstein, and G. Gur-Ari, “The large learning rate phase of deep learning: the catapult mechanism,” *arXiv preprint arXiv:2003.02218*, 2020.

[4] S. Seong, Y. Lee, Y. Kee, D. Han, and J. Kim, “Towards flatter loss surface via nonmonotonic learning rate scheduling.,” in *UAI*, pp. 1020–1030, 2018.

[5] X. Yue, M. Nouiehed, and R. Al Kontar, “Salr: Sharpness-aware learning rate scheduler for improved generalization,” *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[6] L. N. Smith, “A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay,” *arXiv preprint arXiv:1803.09820*, 2018.

[7] L. Chen and J. Bruna, “Beyond the edge of stability via two-step gradient updates,” 2023.

[8] L. E. MacDonald, J. Valmadre, and S. Lucey, “On progressive sharpening, flat minima and generalisation,” *arXiv preprint arXiv:2305.14683*, 2023.

[9] K. Ahn, S. Bubeck, S. Chewi, Y. T. Lee, F. Suarez, and Y. Zhang, “Learning threshold neurons via the" edge of stability",” *arXiv preprint arXiv:2212.07469*, 2022.

[10] A. Damian, E. Nichani, and J. D. Lee, “Self-stabilization: The implicit bias of gradient descent at the edge of stability,” *arXiv preprint arXiv:2209.15594*, 2022.

[11] X. Zhu, Z. Wang, X. Wang, M. Zhou, and R. Ge, “Understanding edge-of-stability training dynamics with a minimalist example,” in *The Eleventh International Conference on Learning Representations*, 2023.

[12] Z. Wang, Z. Li, and J. Li, “Analyzing sharpness along gd trajectory: Progressive sharpening and edge of stability,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 9983–9994, 2022.

[13] I. Kreisler, M. S. Nacson, D. Soudry, and Y. Carmon, “Gradient descent monotonically decreases the sharpness of gradient flow solutions in scalar networks and beyond,” *arXiv preprint arXiv:2305.13064*, 2023.

[14] S. Arora, Z. Li, and A. Panigrahi, “Understanding gradient descent on the edge of stability in deep learning,” in *International Conference on Machine Learning*, pp. 948–1024, PMLR, 2022.

[15] K. Ahn, J. Zhang, and S. Sra, “Understanding the unstable convergence of gradient descent,” in *International Conference on Machine Learning*, pp. 247–257, PMLR, 2022.

[16] K. Lyu, Z. Li, and S. Arora, “Understanding the generalization benefit of normalization layers: Sharpness reduction,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 34689–34708, 2022.

[17] Y. Wang, Z. Xu, T. Zhao, and M. Tao, “Good regularity creates large learning rate implicit biases: edge of stability, balancing, and catapult,” *arXiv preprint arXiv:2310.17087*, 2023.

[18] S. S. Du, W. Hu, and J. D. Lee, “Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced,” in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018.

[19] T. Ye and S. S. Du, “Global convergence of gradient descent for asymmetric low-rank matrix factorization,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 1429–1439, 2021.

[20] C. Ma, Y. Li, and Y. Chi, “Beyond procrustes: Balancing-free gradient descent for asymmetric low-rank matrix sensing,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 867–877, 2021.

A Necessary condition of convergence from stability theory

For an objective function $f(u)$, consider the GD update in terms of a iterative map ψ

$$u_{k+1} = \psi(u_k) := u_k - h\nabla f(u_k).$$

We further consider a stationary point u^* of the objective function f, i.e., $\nabla f(u^*) = 0$. Then this point u^* is a fixed point of the map ψ since

$$u^* = u^* - h\nabla f(u^*) = \psi(u^*).$$

If all the magnitudes of the eigenvalues of Jacobian matrix $\nabla\psi(u^*)$ are less than 1, u^* is a stable fixed point (see more explanation in Section 5 of [2]). Consequently, we have the following theorem

Theorem 3 (Necessary condition)**.** *Let u^* be a local minimum point of $f(u)$ and consider GD updates. If $-I \prec I - h\nabla^2 f(u^*) \prec I$, i.e., $h < \frac{2}{L^*}$, where $\nabla^2 f(u^*) \preceq L^* I$, we have that u^* is a stable fixed point of GD map.*

Note the above theorem is a necessary condition of the convergence of GD to a minimizer. This is due to the fact that if $h > \frac{2}{L^*}$, there exists at least one eigendirection s.t. the magnitude of its eigenvalue is greater than 1. Namely, there will be an unstable direction of the map that prevents GD from converging towards the point u^*.

B Preparation for proofs

Before the proofs, we first take a closer look at the GD iteration for the function $f(x,y) = F(xy)$

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ y_k \end{pmatrix} - h\ell_k \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_k \\ y_k \end{pmatrix} = \begin{pmatrix} 1 & -h\ell_k \\ -h\ell_k & 1 \end{pmatrix} \begin{pmatrix} x_k \\ y_k \end{pmatrix},$$

where $\ell_k = F'(x_k y_k)$. Let $u_k = \begin{pmatrix} x_k \\ y_k \end{pmatrix}$. Then

$$u_{k+1}^\top u_{k+1} = (1 + h^2\ell_k^2) u_k^\top u_k - 4h\ell_k x_k y_k, \tag{4}$$

$$\begin{aligned} u_{k+2}^\top u_{k+2} &= (1 + h^2\ell_{k+1}^2) u_{k+1}^\top u_{k+1} - 4h\ell_{k+1} x_{k+1} y_{k+1} \\ &= (1 + h^2\ell_{k+1}^2)((1 + h^2\ell_k^2) u_k^\top u_k - 4h\ell_k x_k y_k) - 4h\ell_{k+1} x_{k+1} y_{k+1}. \end{aligned} \tag{5}$$

Lemma 1. *Under the same assumption as Theorem 1, we have*

$$u_{k+1}^\top u_{k+1} \lesssim u_k^\top u_k - 4h\ell_k(x_k y_k - \ell_k)$$

Proof. By Lemma 3, we have

$$u_k^\top u_k \le \frac{4}{h} + \mathcal{O}(h).$$

Then by (4), we have

$$u_{k+1}^\top u_{k+1} \le u_k^\top u_k - 4h\ell_k(x_k y_k - \ell_k) + \mathcal{O}(h^3\ell_k^2)$$

□

For the update of $x_k y_k$, we have

$$x_{k+1} y_{k+1} = (1 + h^2\ell_k^2) x_k y_k - h\ell_k u_k^\top u_k, \tag{6}$$

$$x_{k+1} y_{k+1} - 1 = (x_k y_k - 1)\left(1 - h\frac{\ell_k}{x_k y_k - 1}(u_k^\top u_k - h\ell_k x_k y_k)\right). \tag{7}$$

Let $\delta := xy - 1$ and $\delta_k := x_k y_k - 1$. Then we define the following functions:

$$\begin{aligned} \ell(\delta) &= F'(\delta + 1), \quad \text{and then } \ell_k = \ell(\delta_k) = F'(x_k y_k); \\ q(\delta) &= \frac{\ell(\delta)}{\delta}, \quad \text{and then } q(\delta_k) = \frac{\ell_k}{x_k y_k - 1}; \\ r(u_k, \delta) &= 1 - hq(\delta)(u_k^\top u_k - h\ell(\delta)(\delta + 1)), \quad \text{and then } r_k = r(u_k, \delta_k) = 1 - h\frac{\ell_k}{x_k y_k - 1}(u_k^\top u_k - h\ell_k x_k y_k). \end{aligned}$$

Let

$$C_k = \frac{1 - r_k}{q(\delta_k)}, \quad \text{and then } r_k = 1 - C_k q(\delta_k).$$

From Lemma 1, we also define

$$L(\delta) = \ell(\delta)(\delta + 1 - \ell(\delta)), \text{ and then } L(\delta_k) = \ell_k(x_k y_k - \ell_k).$$

All the above functions also depends on a or b. If not specified, all the properties for these functions used in the proofs are valid for all $0 < a \le 1$ or $b = 2n + 1$ with $n \in \mathbb{N}$.

C Proof of Theorem 1 for function f_{good} (2)

The following propositions show the properties of the functions defined above under this objective. The proof is based on simple analysis and Taylor expansion and thus omitted. If not specified, these properties are independent of a.

Proposition 1. *The function* $\ell(\delta) = \frac{2(e^{\delta}-1)}{e^{\delta}+1}$ *has the following properties:*

- $\ell(\delta) = -\ell(-\delta)$
- $\ell(\delta) > 0$ *for* $\delta > 0$.
- $|\ell(\delta)| \leq 3$ *for all* $\delta > 0$.

Proposition 2. *The function* $L(\delta) = \ell(\delta)(\delta + 1 - \ell(\delta))$ *has the following properties:*

- $L(\delta)$ *monotonically increases for* $\delta \geq 0$, *and* $L(\delta) \geq L(0)$ *for* $\delta \geq 0$.
- $L(\delta) + L(-\delta) \geq 0$ *for all* δ.
- $L(\delta_1) + L(\delta) \geq 0$ *for* $\delta_1 \geq 1$ *and all* δ.
- $L(\delta) + L(r\delta) \geq 0.8(1 + r)\delta$ *for* $\delta > 0$ *and* $-1 < r < 0$.

Proposition 3. *The function* $q(\delta) = \frac{\ell(\delta)}{\delta}$ *has the following properties:*

- $q(\delta)$ *is symmetric with respect to* $\delta = 0$, *i.e.,* $q(\delta) = q(-\delta)$.
- *For* $\delta \geq 0$, $q(\delta)$ *monotonically decreases as* δ *increases, and then* $q(\delta) \leq q(0) = 1$.
- *From Taylor expansion, we have*
$$q(\delta) \geq 1 - c_1\delta^2, \text{ where } c_1 = \frac{1}{12} > 0.$$
Also, when $\delta \leq 1$,
$$q(\delta) \leq 1 - c_2\delta^2, \text{ where } c_2 = \frac{1}{15}.$$
- *For* $2 \leq C \leq 4$,
$$\delta = q^{-1}(\frac{1}{C}) \leq 2C.$$

From the above propositions, we will use $\delta_k = |x_k y_k - 1|$ instead of $x_k y_k - 1$ for the rest of the proofs.

C.1 Proof of convergence

Proof. By Proposition 3 and Proposition 1,
$$r_k = 1 - hq(\delta_k)(u_k^\top u_k - h\ell_k x_k y_k) < 1.$$

As is discussed in Lemma 4, the initial condition set removes a null set of converging initial conditions within finite steps, i.e., $r_k \neq 0$. If $r_k > 0$, then $|x_k y_k - 1|$ will monotonically decreases. Otherwise, $r_k < 0$ and by Lemma 5, for all $n \geq k$, if $x_n y_n > 1$, then $x_{n+1} y_{n+1} < 1$ and vice versa. Then we can consider the kth iteration when $x_k y_k > 1$ and we have $x_{k+2n} > 1$ for $n = 1, \cdots,$.

By Lemma 8, $|x_k y_k - 1|$ is guaranteed to decrease monotonically in $|x_k y_k - 1| \leq R_2$, with $|r_k| < 1$. Therefore, GD will converge to $xy = 1$ (otherwise, if $x_k y_k$ converges to $c \neq 1$, with $|r_k| < 1$, $|x_k y_k - 1|$ will keep decreasing, contradiction). □

C.2 Proofs of EoS

Proof of Part II: limiting sharpness. From the proof above, GD converges to a global minimum. Throughout this proof, we will use the big $\mathcal{O}$ notation for complexity and distance. The proof can be made more rigorous by considering specific constant scaling of these orders.

By the lower bound of h, $u_0^\top u_0 \gtrsim \frac{2/q(1)+\mathcal{O}(h)}{h}$, i.e., $|r_0(1)| \geq 1 + \mathcal{O}(h)$. Then we consider the decrease of $u_k^\top u_k$ until GD enter the region $|r_k| < 1$. More precisely, we consider three regions: 1) when $|\delta_k|$ is small enough s.t. $u_k^\top u_k$ does not decrease every two steps. By the following bound,

$$\begin{aligned}
u_{k+2}^\top u_{k+2} &= (1 + h^2\ell_{k+1}^2)u_{k+1}^\top u_{k+1} - 4h\ell_{k+1}x_{k+1}y_{k+1}\\
&= (1 + h^2\ell_{k+1}^2)((1 + h^2\ell_k^2)u_k^\top u_k - 4h\ell_k x_k y_k) - 4h\ell_{k+1}x_{k+1}y_{k+1}\\
&\geq u_k^\top u_k - 4h[\ell_k x_k y_k - \frac{1}{2}\ell_k^2 + \ell_{k+1}x_{k+1}y_{k+1} - \frac{1}{2}\ell_{k+1}^2] - 4h^3\ell_{k+1}^2\ell_k x_k y_k + h^4\ell_{k+1}^2\ell_k^2 u_k^\top u_k\\
&\geq u_k^\top u_k - 4h[(1 + r_k)(x_k y_k - 1) + \frac{1}{2}(1 + r_k^2)(x_k y_k - 1)^2] - 4h^3\ell_{k+1}^2\ell_k x_k y_k + h^4\ell_{k+1}^2\ell_k^2 u_k^\top u_k,
\end{aligned}$$

we have that such $|\delta_k| \leq \mathcal{O}(h)$. 2) Starting from $|\delta_k| = \mathcal{O}(h)$, consider $|\delta_k|$ increases to some region that is $\mathcal{O}(h)$ away from 1, i.e., the rate of increase is at least $1 + \mathcal{O}(h)$. In this region, it takes GD at most $\mathcal{O}(-\log h/h)$ steps and $u_k^\top u_k$ may decreases every two steps, where the order of decrease is $\mathcal{O}(h)$. 3) Starting from the end of 2), if the rate of increase is at least $1 + \mathcal{O}(h)$, then the complexity of entering the region $|r_k| < 1$ is at most $\mathcal{O}(1/h)$ which follows the same derivation as 2). Otherwise, i.e., the rate of increase is less than $1 + \mathcal{O}(h)$. Since the decrease of $u_k^\top u_k$ is $\mathcal{O}(h)$ and $q(\delta_k)$ keep decreasing before $|r_k| < 1$, it takes GD at most $\mathcal{O}(1/h)$ steps such that $|r_k|$ is $\mathcal{O}(h)$ less than $1 + \mathcal{O}(h)$, i.e., $|r_k| < 1$. Therefore, the overall decrease of $u_k^\top u_k$ is at most $\tilde{\mathcal{O}}(1)$ and we still have $u_k^\top u_k \geq \frac{2/q(1) - \tilde{\mathcal{O}(h)}}{h}$ at the end of all the above processes. Also, according to Lemma 6, $u_k^\top u_k$ will decrease to at least $\frac{2/q(1)+\mathcal{O}(h)}{h}$. Otherwise, $|r_k| > 1$ in $|x_k y_k - 1| \geq R_1$ and therefore $u_k^\top u_k$ will keep decreasing.

Before further discussing the complexity of GD, we still need an upper bound of $x_k y_k$. Let $u_k^\top u_k = \frac{C_k}{h}$. Then since $r_k < 0$ for all k, we can consider the maximum δ s.t. $r_0 \approx 0$, i.e., $\delta = q^{-1}(\frac{1}{C_k}) \leq 2C_k$ by Proposition 3. Therefore, $\delta_k \lesssim 2C_k \leq 5$. Here we just relax C_k to be upper bounded by 2.5 based on $2/q(1) + \mathcal{O}(h)$.

If $|x_k y_k - 1| > R_2$, suppose $|x_k y_k - 1| > 1.5$ then $|r_k| < r_k(1.5) < 0.8 + \mathcal{O}(h^2) < 0.85$. Then the complexity of entering $\delta_k \leq 1.5$ is $\frac{\log(1.5/5)}{\log(0.85)} = \mathcal{O}(1)$. Therefore, the step of GD entering this region is less than the order $\mathcal{O}(1/\log(r_k(1.5))) = \tilde{\mathcal{O}}(1)$.

Then consider the complexity of GD entering $\{r < 1, |xy - 1| < 1 - \mathcal{O}(h^2)\}$ (By Lemma 7, $R_2 \geq 1 - \mathcal{O}(h^2)$). We can consider $|r_k|$ via the map $g(\delta) = \delta(C_k q(\delta) - 1)$, where we ignore the $\mathcal{O}(h^2)$ terms and fix $C_k = C_N$ for some N. We would like to analyze the complexity of the map converging to $\mathcal{O}(h)$ error of the fixed point δ^* near $\delta = 1$, i.e., $(C_k q(\delta) - 1) = 1$. Since $(C_k q(1 - \mathcal{O}(h^2)) - 1) = 1 - \mathcal{O}(h)$,

$$\begin{aligned}
1 = C_k q(\delta^*) - 1 &\geq C_k(1 - c_1(\delta^*)^2) - 1\\
&= C_k(1 - c_1) - 1 + (1 - (\delta^*)^2)C_k c_1\\
&= 1 - \mathcal{O}(h) + (1 - (\delta^*)^2)C_k c_1
\end{aligned}$$

Therefore $\delta^* \geq 1 - \mathcal{O}(h/c_1)$.

By checking the derivative of the map, we have $g'(\delta) = C_k q(\delta) - 1 + C_k \delta q'(\delta)$ and $q'(\delta) < -c_1$ for $0.8 < \delta < 1.5$. Since $C_k > 2$, when $\delta > \delta^*$, $C_k q(\delta) - 1 < 1$ and thus $0 < g'(\delta) \leq 1 - C_k \delta^* c_1$. Then the map will decrease from above to enter the region within $\mathcal{O}(h)$ distance of δ^* with the complexity of $\mathcal{O}(-\log(h)/c_1)$. For the $\mathcal{O}(h^2)$ terms, the error between this map and the true δ_k update is of the order $\mathcal{O}(\exp(h^2 \cdot \log(h)/c_1) - 1) < \mathcal{O}(h)$ and thus can be omitted. Then GD starts to decrease in the monotone decreasing region of $|x_k y_k - 1|$ with $u_k^\top u_k \geq \frac{2}{h} + \mathcal{O}(1)$.

Next we consider the Nth iteration with $u_N^\top u_N = \frac{2}{h} + \bar{C}_N$, for some $\bar{C}_N = \mathcal{O}(1)$. For $k \geq N$, $u_k^\top u_k > \frac{2}{h}$, and $x_k y_k > 1$, we have that $\bar{C}_k - \bar{C}_{k+2} = \mathcal{O}(h)$. We would like to prove that eventually $|r_k| \geq 1 - \mathcal{O}(h^2)$ when $u_k^\top u_k \geq \frac{2}{h}$, i.e., we would like to analyze the complexity of $|r_k|$ increasing to $|r_k| = 1 - \mathcal{O}(h^2)$.

First note

$$|r_k| \geq q(\delta_k)h(u_k^\top u_k - h\ell_k x_k y_k) - 1 = 1 - 2c_1\delta_k^2 + 2\bar{C}_N h \pm \mathcal{O}(h^2).$$

Then the complexity is upper bounded by the complexity of δ_k decreasing to the value s.t. $-2c_1\delta_k^2 + 2\bar{C}_N h = 0$, i.e., $\delta_k = \mathcal{O}(\sqrt{h/c_1})$. Let $c_1 = h^p$. By the assumption of h, we have $p \leq \frac{3}{4}$. Therefore $\sqrt{h/c_1} = h^{\frac{1-p}{2}}$. Then we analyze

$$\delta_{n+N} = \delta_N - 2h^p\delta_N^3 + 2\bar{C}_N h\delta_N - 2h^p\delta_N^3(1 - 2h^p\delta_N^2 + 2\bar{C}_N h)^2 + \cdots = \mathcal{O}(h^{\frac{1-p}{2}})$$

We further remove all the $\bar{C}_N$ terms since the sum of them is $\mathcal{O}(h) < \mathcal{O}(h^{\frac{1-p}{2}})$. We first consider

$$\delta_N - 2h^p\delta_N^3 - 2h^p\delta_N^3(1 - 2h^p\delta_N^2)^2 + \cdots = h^{p_1}$$

Ignoring the $\mathcal{O}(h^2)$ part, we have that $|r_k|$ is increasing as $|\delta_k|$ decreases. Then

$$h^{p_1} = LHS \lesssim \delta_N - h^p h^{3p_1} - h^p h^{3p_1} + \cdots = \delta_N - n_1 h^{3p_1+p} \le 1 - n_1 h^{3p_1+p}$$

Then

$$n_1 \le h^{-(3p_1+p)}$$

Iteratively, we consider $\delta_N = h^{p_1}$ and $p_2 = \frac{4}{3}p_1$. Then

$$h^{p_2} = LHS \lesssim h^{p_1} - h^p h^{3p_2} - h^p h^{3p_2} + \cdots = \delta_N - n_2 h^{3p_2+p} \le h^{p_1} - n_2 h^{3p_2+p}$$

$$n_2 \le h^{-(3p_2+p-p_1)} = h^{-(3p_1+p)}$$

We can next solve i s.t.

$$\left(\frac{4}{3}\right)^{i-1} p_1 = \frac{1-p}{2} \Rightarrow i = 1 + \frac{\log\frac{1-p}{2p_1}}{\log\frac{4}{3}}$$

From the above, we also need to consider the $\bar{C}_N$ part and thus require

$$3p_1 + p < \min\{\frac{1+3p}{2}, 1\}$$

Then

$$\frac{1-p}{2p_1} > \frac{3}{2} \text{ and } i \text{ is at most } \mathcal{O}(\log h)$$

Then the total complexity of achieving $h^{\frac{1-p}{2}}$ is $\mathcal{O}(h^{-(3p_1+p)}) < \mathcal{O}(h^{-1})$. Thus we will eventually have $|r_k| = 1 - \mathcal{O}(h^2)$ when $u_k^\top u_k \ge \frac{2}{h}$. For the $k+2$th iteration, the increase of $1 - c\delta_k^2$ is $\mathcal{O}((1 - |r_k|)\delta_k^2 + (1 - |r_{k+1}|)\delta_k^2) = \mathcal{O}((1 - |r_k|)\delta_k^2)$, where one step change of $|r_k|$ is at most $\mathcal{O}(h^2)$ and thus can be omitted, and the decrease of $h(u_k^\top u_k - h\ell_k x_k y_k)$ is $\mathcal{O}(h^2\delta_k(1 - |r_k|))$ by Lemma 6. Thus when $\delta_k \ge \mathcal{O}(h^2)$, $\mathcal{O}((1-|r_k|)\delta_k^2) \ge \mathcal{O}(h^2\delta_k(1-|r_k|))$, i.e., $|r_k| \ge 1-\mathcal{O}(h^2)$; when $\delta_k < \mathcal{O}(h^2)$, from the expression of $|r_k|$, we have $|r_k| = 1 - \mathcal{O}(h^2)$ for the kth iteration s.t. $u_k^\top u_k \ge \frac{2}{h} - \mathcal{O}(h)$. Therefore, $|r_k| \ge 1 - \mathcal{O}(h^2)$ for all the kth iteration s.t. $u_k^\top u_k \ge \frac{2}{h} - \mathcal{O}(h)$.

Next, consider the first step when $\frac{2}{h} \ge u_N^\top u_N \ge \frac{2}{h} - \mathcal{O}(h)$ and $x_N y_N > 1$ since the decrease of $u_k^\top u_k$ at each step is at most $\mathcal{O}(h)$. Then based on the expression of r_N, we have $|r_N| \le 1 - \mathcal{O}(h^2)$ and additionally by Lemma 6, we have $|r_k| = 1 - \mathcal{O}(h^2)$ for all $k \ge N$. By series expansion, $\ell_{N+k}x_{N+k}y_{N+k} = \mathcal{O}(|x_{N+k}y_{N+k} - 1|)$. Then

$$\begin{aligned}
{u_{N+k+2}}^\top u_{N+k+2} &\ge {u_{N+k}}^\top u_{N+k} - 4h[\ell_{N+k}x_{N+k}y_{N+k} + \ell_{N+k+1}x_{N+k+1}y_{N+k+1}] - \mathcal{O}(h^3) \\
&\ge {u_{N+k}}^\top u_{N+k} - \mathcal{O}\left(h[\ell_{N+k}x_{N+k}y_{N+k} + \ell_{N+k+1}x_{N+k+1}y_{N+k+1}]\right) \\
&\ge {u_{N+k}}^\top u_{N+k} - \mathcal{O}\left(h[(1 - |r_{N+k}|)\delta_{N+k}]\right) \\
&\ge {u_N}^\top u_N - \mathcal{O}\left(h\sum_{i=0}^{k}[(1 - |r_{N+i}|)\delta_{N+i}]\right) \\
&\ge {u_N}^\top u_N - \mathcal{O}\left(h\sum_{i=0}^{k}[(1 - |r_{N+i}|)\prod_{j=0}^{i} r_{N+j}\delta_N]\right) \\
&\ge {u_N}^\top u_N - \mathcal{O}\left(h^3\delta_N\sum_{i=0}^{k}(1 - \mathcal{O}(h^2))^i\right)
\end{aligned}$$

Take $k \to \infty$ for both side and we have $u_\infty^\top u_\infty \ge u_N^\top u_N - \mathcal{O}(h)$. □

Proof of Part I: end of pre-EoS and preparation for progressive sharpening. Let $a = 1$ for the rest of the proof. Consider the Hessian $\nabla^2 f$. The trace is

$$\mathrm{tr}(\nabla^2 f) = (x^2 + y^2)G_1(xy - 1),$$

where

$$\begin{aligned}
G_1(\delta) &= \frac{2^{2-a}\log^{1-a}(2)\left(\log\left(e^{-\delta}+1\right)+\log\left(e^{\delta}+1\right)\right)^{a-2}\left((a-1)\left(e^{\delta}-1\right)^2+2e^{\delta}\log\left(e^{-\delta}+1\right)+2e^{\delta}\log\left(e^{\delta}+1\right)\right)}{\left(e^{\delta}+1\right)^2}\\
&= \frac{2^{3-a}e^{\delta}\log^{1-a}(2)\left(\log\left(e^{-\delta}+1\right)+\log\left(e^{\delta}+1\right)\right)^{a-1}}{\left(e^{\delta}+1\right)^2}\\
&+ \frac{2^{2-a}(a-1)\left(e^{\delta}-1\right)^2\log^{1-a}(2)\left(\log\left(e^{-\delta}+1\right)+\log\left(e^{\delta}+1\right)\right)^{a-2}}{\left(e^{\delta}+1\right)^2}\\
&= q(\delta)\frac{2\delta}{e^{\delta}-e^{-\delta}}+q(\delta)\frac{(a-1)\left(e^{\delta}-1\right)\delta}{\left(e^{\delta}+1\right)\left(\log\left(e^{-\delta}+1\right)+\log\left(e^{\delta}+1\right)\right)}\\
&\le q(\delta)\frac{2\delta}{e^{\delta}-e^{-\delta}}+q(\delta)(a-1)\left(1-\frac{2\delta}{e^{\delta}-e^{-\delta}}\right)\\
&= \left((2-a)\frac{2\delta}{e^{\delta}-e^{-\delta}}-(1-a)\right)q(\delta).
\end{aligned}$$

Let

$$q_1(\delta,a)=(2-a)\frac{2\delta}{e^{\delta}-e^{-\delta}}-(1-a).$$

Then both $q_1(\delta,a)$ and $q(\delta)$ decreases as δ increases for $\delta \ge 0$. Also, fix δ, $q_1(\delta,a)$ and $q(\delta,a)$ increases as a increases.

The determinant is

$$\det(\nabla^2 f)=G_2(xy-1)$$

where

$$\begin{aligned}
G_2(\delta) = -&\frac{4^{2-a}\left(e^{\delta}-1\right)\log^{2-2a}(2)\left(\log\left(e^{-\delta}+1\right)+\log\left(e^{\delta}+1\right)\right)^{2a-3}}{\left(e^{\delta}+1\right)^3}\\
&\times\left(2(a-1)(\delta+1)\left(e^{\delta}-1\right)^2+\left(4e^{\delta}(\delta+1)+e^{2\delta}-1\right)\left(\log\left(e^{-\delta}+1\right)+\log\left(e^{\delta}+1\right)\right)\right).
\end{aligned}$$

Then

$$G_2(\delta)\sim\mathcal{O}(\delta^{2a-2}) \text{ as } \delta\to\infty, \text{ and } G_2(0)=0.$$

Thus we have that $G_2(\delta)$ is bounded since $G_2(\delta)\in\mathcal{C}^1$. Therefore, by Taylor expansion, the largest eigenvalue of Hessian (sharpness) is upper bounded by

$$G_1(\delta)(x^2+y^2)+\mathcal{O}\left(\frac{|G_2(xy-1)|}{x^2+y^2}\right),$$

where $|G_2(xy-1)|$ is bounded due to the boundedness of x_0y_0 in the initial condition set and the convergence of the GD trajectory (see more details in the proof of limiting sharpness).

Let $h=\frac{C}{x_0^2+y_0^2}$. If $|r_0|<1$, then

$$q(\delta_0)\lesssim\frac{2}{C}.$$

Otherwise, similar to the proof of part II, $u_k^\top u_k$ decreases at most $\tilde{\mathcal{O}}(1)$ before GD enters the region where δ_k starts to decrease for the first time, i.e., $|r_k|<1$. Therefore, for such $\delta=\delta_k$, we also have

$$1-c_1\delta^2\le q(\delta)\lesssim\frac{2}{C},$$

where $\tilde{\mathcal{O}}(h)$ term is omitted. Then

$$\delta\gtrsim\frac{\sqrt{C-2}}{\sqrt{Cc_1}},$$

and

$$G_1(\delta)\lesssim\frac{2}{C}q_1(\frac{\sqrt{C-2}}{\sqrt{Cc_1}},a)\le\frac{2}{C}q_1(\frac{\sqrt{C-2}}{\sqrt{C/12}},1),$$

where the last inequality follows easily from checking the derivative of $q_1(\frac{\sqrt{C-2}}{\sqrt{Cc_1(a)}},a)$ w.r.t. a. Also, $S_\infty\approx\frac{2}{h}=\frac{2}{C}u_0^\top u_0$. Based on the above discussion, $u_k^\top u_k\le u_0^\top u_0-\mathcal{O}(h^{1-m})$. Therefore,

$$S_k\lesssim G_1(\delta_k)u_k^\top u_k\lesssim q_1(\frac{\sqrt{C-2}}{\sqrt{C/12}},1)S_\infty\le\frac{1}{4}(6-C)S_\infty,$$

where the last inequality follows from linearization and shift. □

C.3 Proof of balancing

Proof. By the limiting sharpness in Theorem 1 and the global minima $xy = 1$,

$$(x_\infty - y_\infty)^2 = u_\infty^\top u_\infty - 2 \le \frac{2}{h} - 2 = \frac{2}{C}(x_0^2 + y_0^2 - 2x_0y_0) + \frac{4}{C}x_0y_0 - 2 \le \frac{2}{C}(x_0 - y_0)^2 + 2(x_0y_0 - 1).$$

□

C.4 Supplementary lemmas

We first show a summary of the behavior of $u_k^\top u_k$ in the following lemma, which is the main idea of the proof of convergence.

Lemma 2. *Under the assumption of Theorem 1, there exist an increasing sequence $\{2N_i\}_{i\in\mathbb{Z}}$ with $N_i \in \mathbb{Z}$ and $N_{i+1} > N_i$, s.t., $u_{2N_{i+1}}{}^\top u_{2N_{i+1}} \le u_{2N_i}{}^\top u_{2N_i}$. More previsely, consider the even number of iteration, i.e., $k = 2i$ for $i \in \mathbb{Z}$,*

- *Stage I (not necessary): $u_k^\top u_k$ increases but is bounded by $\frac{4}{h} + \mathcal{O}(h)$. This happens when $|x_ky_k - 1|$ is too small at the early stage of iterations (see Lemma 3);*
- *Stage II: $u_k^\top u_k$ decreases every two steps when x_k, y_k is outside a 'monotone decreasing region' (see Lemma 6);*
- *Stage III: when x_k, y_k is near the 'monotone decreasing region', there exists N, s.t. $u_{k+2N}^\top u_{k+2N}$ is smaller than $u_k^\top u_k$ (see Lemma 8);*
- *Stage IV: when $|x_ky_k - 1|$ monotonically decreases, $u_k^\top u_k$ decreases every two steps (see Lemma 6 and Lemma 8)*

Lemma 3. *Under the assumption of Theorem 1, we have*

$$u_k^\top u_k \le \frac{4}{h} + \mathcal{O}(h), \forall k \in \mathbb{N}.$$

Proof. By Lemma 5 and Lemma 6, we know that the increase of $u_k^\top u_k$ after two step can only happen when $r_k \le -1$ and $0 < x_ky_k - 1 < 1$.

WOLG assume N is the first step s.t. $u_N^\top u_N = 4/h + \mathcal{O}(h)$. By Lemma 6, we would like to show that there exists k s.t. for $x_ky_k > 1$, we have either $x_ky_k - 1 > 1$ or $-1 < r_k < 0$.

Let $\delta_k = |x_ky_k - 1|$. Then

$$|r_k| \ge 4(1 - c_1\delta_k^2) - 1 + \mathcal{O}(h^2) \ge 4(1 - \delta_k^2/4) - 1 + \mathcal{O}(h^2)$$

and therefore

$$\delta_{k+1} \ge \delta_k(3 - 4c_1\delta_k^2 + \mathcal{O}(h^2)) \ge \delta_k(3 - \delta_k^2 + \mathcal{O}(h^2)).$$

If $\delta_k < 1$, it takes GD $\mathcal{O}(-\log(\delta_0))$ steps to go to $\delta_k \ge 1$. When $\delta_k \ge 1$ $u_k^\top u_k$ already starts to decrease every two steps. Moreover,

$$\delta_{k+1} \ge \delta_k(3 - \delta_k^2 + \mathcal{O}(h^2)) = \delta_k + 2\delta_k - \delta_k^3 + \mathcal{O}(h^2).$$

Therefore, δ_k keeps increasing for a while, staying in $\delta_k \ge 1$. Therefore, $u_k^\top u_k$ stays in the decreasing region with increase of at most $\mathcal{O}(h)$ for each step (for more detailed version, see Lemma 2), i.e., $u_k^\top u_k \le \frac{4}{h} + \mathcal{O}(h)$ for all k. □

Lemma 4. *Under the assumption of Theorem 1, GD does not converge outside $\{(x,y)|x^2+y^2 \le 2/h\}$.*

Proof. We first would like to remove the initial condition that can converge in finite steps to the minima outside of this region. It turns out that such initial conditions form a null set. The proof is almost the same as [2] except for some easy calculations and thus omitted. We further remove all the initial conditions that converges to the periodic orbits, i.e., $\prod_{i=1}^n r_k \cdots r_{k+2i-1} = 1$ for all k and n. By similar argument in [2] (i.e., for each k, n, this is a null set, and therefore the union of countably many null sets is still a null set), such initial conditions also form a null set.

Assume $u_k^\top u_k \ge 2/h + \epsilon_0$ for all k, where $\epsilon_0 > 0$. Consider the case when $|x_ky_k - 1| < \epsilon_1 = \frac{\sqrt{h\epsilon_0}}{2}$. Then

$$\begin{aligned} |r_k| &\ge |h(1 - \mathcal{O}(\epsilon_1^2))(2/h + \epsilon_0) + \mathcal{O}(h^2\epsilon_1) - 1| \\ &= |1 - 2\epsilon_1^2 + h\epsilon_0 + \mathcal{O}(h^2\epsilon_1)| \\ &= |1 + \frac{h}{2}\epsilon_0| > 0 \end{aligned}$$

Namely, GD cannot converge with $u_k^\top u_k > 2/h$. □

Lemma 5. *Under the assumption of Theorem 1, if $r_k < 0$, then $r_n < 0$ for all $n \geq k$.*

Proof. From Lemma 1, we have $|C_{k-1} - C_k| = \mathcal{O}(h^2)$. Also $C_k = \frac{1-r_k}{q(\delta_k)}$. If $r_k < 0$,

$$\begin{aligned} r_{k+1} = 1 - C_{k+1}q(\delta_{k+1}) &\leq 1 - \frac{1-r_k}{q(\delta_k)}q(\delta_{k+1}) + \mathcal{O}(h^2) \\ &= 1 - \frac{1-r_k}{q(\delta_k)}q(r_k\delta_k) + \mathcal{O}(h^2) \end{aligned}$$

By Proposition 3, it suffices to prove that

$$R_0(\delta_k, -r_k) := (1-r_k)q(-r_k\delta_k) - q(\delta_k) > 0, \ \forall \delta_k > 0$$

We abuse the notation and let $r_k(\delta) = 1 - C_kq(\delta)$ for fixed $C_k > 0$. By Proposition 3, $r_k(\delta)$ monotonically increases for $\delta > 0$ and there is only one root denoted as δ_0 s.t. $r_k(\delta) = 0$ (Since $r_k < 0$, there exists δ s.t. $r_k(\delta) < 0$ and when δ is large, $r_k(\delta) > 0$; therefore, it has a root). When $r_k(\delta_0) = r_k = 0$, by Proposition 3, $R_0(\delta_0, -r_k(\delta_0)) = R_0(\delta_0, 0) = 1 - q(\delta_0) > 0$. Also

$$\begin{aligned} R_0(\delta, -r_k(\delta)) &= C_kq(\delta)q\left((C_kq(\delta)-1)\delta\right) - q(\delta) \\ &= q(\delta)\left[C_kq\left((C_kq(\delta)-1)\delta\right) - 1\right] \end{aligned}$$

Therefore, we only need to show $K(\delta) := (C_kq(\delta)-1)\delta < \delta_0$ for $\delta \in (0, \delta_0)$. The derivative of K is

$$K'(\delta) = C_kq(\delta) - 1 + C_k\delta q'(\delta) < C_kq(\delta) - 1.$$

Then the maximum point of $K(\delta)$ is achieved at $\delta_1 < \delta_0$ by Proposition 3. Thus, $R_0(\delta, -r_k(\delta)) > 0$, and consequently

$$R_0(\delta_k, -r_k) \geq q(\delta_k)(C_kq(K(\delta_1)) - 1) > 0,$$

$$r_{k+1} \leq -\frac{R_0(\delta_k, -r_k)}{q(\delta_k)} + \mathcal{O}(h^2) \leq -(C_kq(R_1(\delta_1)) - 1) + \mathcal{O}(h^2) < 0,$$

since this $\mathcal{O}(h^2)$ is indeed bounded by $c\,\delta_kh^2$ for some universal constant $c > 0$ and can be controlled by the first term. □

Lemma 6. *Under the assumption of Theorem 1, the following properties are the two step decrease of $u_k^\top u_k$ in different cases:*

1. *If $r_k < 0$ and $u_k^\top u_k \leq \frac{4}{h} + \mathcal{O}(h)$, there exists $R_1(a, r_k) \leq 1$, s.t., when $x_ky_k - 1 > R_1(a, r_k)$, we have*
$$u_{k+2}^\top u_{k+2} \leq u_k^\top u_k - \frac{1}{2}h + \mathcal{O}(h^3\ell_{k+1}^2).$$
2. *If $-1 < r_k < 0$ and $0 < x_ky_k - 1 \leq 1$, then*
$$u_{k+2}^\top u_{k+2} \leq u_k^\top u_k - 3.2h(1+r_k)(x_ky_k - 1).$$
3. *If $r_k \leq -1$, $u_k^\top u_k \leq \frac{3}{h}$, and $x_ky_k < 1$, then*
$$u_{k+2}^\top u_{k+2} \leq u_k^\top u_k - \min\{\mathcal{O}(h), \mathcal{O}(h(x_ky_k - 1))\}.$$

Proof. First we consider the two step update of $u_k^\top u_k$ in the following

$$\begin{aligned} u_{k+2}^\top u_{k+2} &= (1 + h^2\ell_{k+1}^2)u_{k+1}^\top u_{k+1} - 4h\ell_{k+1}x_{k+1}y_{k+1} \\ &= (1 + h^2\ell_{k+1}^2)((1 + h^2\ell_k^2)u_k^\top u_k - 4h\ell_kx_ky_k) - 4h\ell_{k+1}x_{k+1}y_{k+1} \\ &\leq u_k^\top u_k - 4h[\ell_kx_ky_k - \ell_k^2 + \ell_{k+1}x_{k+1}y_{k+1} - \ell_{k+1}^2] - 4h^3\ell_{k+1}^2\ell_kx_ky_k + h^4\ell_{k+1}^2\ell_k^2u_k^\top u_k + \mathcal{O}(h^3\ell_{k+1}^2) \\ &= u_k^\top u_k - 4h[\ell_kx_ky_k - \ell_k^2 + \ell_{k+1}x_{k+1}y_{k+1} - \ell_{k+1}^2] + \mathcal{O}(h^3\ell_{k+1}^2). \end{aligned}$$

It surffices to analyze

$$\ell_kx_ky_k - \ell_k^2 + \ell_{k+1}x_{k+1}y_{k+1} - \ell_{k+1}^2 = L(\delta_k) + L(r_k\delta_k), \text{ where } L(\delta) = \ell(\delta)(\delta + 1 - \ell(\delta)).$$

By checking the slop and values of the above function, we have

$$L(\delta_k) + L(r_k\delta_k) \geq L(1) + L(r_k) > 0.14$$

for all $r_k < 0$, $0 < a \leq 1$, and $\delta_k \geq 1$.

When $0 < x_ky_k - 1 \leq 1$, in order to make $-1 < r_k < 0$, we should at least have $u_k^\top u_k \leq \frac{4}{h}$. Also we have $\ell_k \leq x_ky_k - 1 \leq 1$ in this region by Proposition 3. Therefore, by Proposition 2

$$\begin{aligned} u_{k+2}^\top u_{k+2} &= (1 + h^2\ell_{k+1}^2)u_{k+1}^\top u_{k+1} - 4h\ell_{k+1}x_{k+1}y_{k+1} \\ &= (1 + h^2\ell_{k+1}^2)((1 + h^2\ell_k^2)u_k^\top u_k - 4h\ell_kx_ky_k) - 4h\ell_{k+1}x_{k+1}y_{k+1} \\ &\leq u_k^\top u_k - 4h[\ell_kx_ky_k - \ell_k^2 + \ell_{k+1}x_{k+1}y_{k+1} - \ell_{k+1}^2] - 4h^3\ell_{k+1}^2\ell_kx_ky_k + h^4\ell_{k+1}^2\ell_k^2u_k^\top u_k \\ &\leq u_k^\top u_k - 4h[\ell_kx_ky_k - \ell_k^2 + \ell_{k+1}x_{k+1}y_{k+1} - \ell_{k+1}^2] \\ &\leq u_k^\top u_k - 3.2h(1 + r_k)(x_ky_k - 1). \end{aligned}$$

When $r_k \le -1$, $u_k^\top u_k \le \frac{3}{h}$, and $x_k y_k < 1$, by Taylor series and simple calculation, we have $L(\delta) + L(r\delta) \ge \min\{0.5(1+r)\delta, L(-1) + L(1)\}$, where $L(-1) + L(1) \ge 0.14$. Then

$$\begin{aligned} u_{k+2}^\top u_{k+2} &\le u_k^\top u_k - 4h[\ell_k x_k y_k - \ell_k^2 + \ell_{k+1} x_{k+1} y_{k+1} - \ell_{k+1}^2] - h(\ell_k^2 + \ell_{k+1}^2) + \mathcal{O}(h^3 \ell_{k+1}^2) \\ &\le u_k^\top u_k - 4h \min\{0.5(1+r_k)(x_k y_k - 1), L(-1) + L(1)\}. \end{aligned}$$

□

Lemma 7. *Under the assumption of Theorem 1, if $-1 \le r_k < 0$, there exists $R_2(a, r_k) \ge 1 - \mathcal{O}(h^2)$, s.t., for $|x_k y_k - 1| \le R_2(a, r_k)$, we have $r_{k+1} > -1$.*

Proof. Since $r_k = 1 - C_k q(\delta_k)$, we have

$$r_{k+1} = 1 - C_{k+1} q(\delta_{k+1}) = 1 - \frac{1 - r_k}{q(\delta_k)} q(r_k \delta_k) \pm \mathcal{O}(h^2).$$

We denote

$$\delta r(r_k, \delta_k) := 2 - \frac{1 - r_k}{q(\delta_k)} q(r_k \delta_k).$$

When $\delta_k > 0$, $\delta r(r_k, \delta_k)$ monotonically decreases as δ_k increases. Consider

$$\delta r(r_k, 1) = 2 - \frac{1 - r_k}{q(1)} q(r_k).$$

This function $\delta r(r_k, 1)$ monotonically decreases when r_k increases between -1 and 0 and $\delta r(-1, 0) > 0$. Therefore, $r_{k+1} > -1 \pm \mathcal{O}(h^2)$ and the conclusion follows from series expansion. □

Lemma 8. *Under the assumption of Theorem 1, if $r_k < 0$ for all $k \ge n$ given some $n \ge 0$, $|x_k y_k - 1|$ will eventually start to decrease in the region $|x_k y_k - 1| < R_2$.*

Proof. By Lemma 6, when $|x_k y_k - 1| \ge R_1(a, r_k)$, $u_k^\top u_k$ keeps decreasing every two step with certain amount away from 0. Eventually, by the expression of $|r_k|$, $u_k^\top u_k$ will decrease until GD enters $|x_k y_k - 1| \le R_2(a, r_k)$. Note it can be checked that $R_1 \le 1 < 1.5 \le R_2$. However, $|x_k y_k - 1|$ may not keep decreasing inside this region. WOLG, consider a lower bound of $R_2(a, r_k)$ to be $R_2 \ge 1 - \mathcal{O}(h^2)$ and we will use R_2 independent of iterations. From the expression of $|r_k|$, we have at least $u_k^\top u_k \le \frac{3}{h}$ in this case.

First consider $|x_k y_k - 1| > R_2$. According to Lemma 6 and Proposition 3, $u_k^\top u_k$ decreases every two steps and the function q decreases when $|x_k y_k - 1|$ increases. Therefore there exists k s.t. $|r_k| < 1$. Moreover, for $|r_k| < 1$, there exists n, s.t. $x_{k+2n} y_{k+2n} < x_k y_k$ and $|r_{k+2n}| < 1$ if $x_{k+2n} y_{k+2n} - 1 \ge R_1$.

Next, assume k is such that $|x_k y_k - 1| > R_2$ but $|x_{k+2} y_{k+2} - 1| \le R_2$. According to the initial condition, there is no periodic orbit in the trajectory (see details in the proof of Lemma 4). We claim that there exists n, s.t., $|r_{k+2n}| < 1$ and $|x_{k+2n} y_{k+2n} - 1| \le R_2$. Otherwise, if $u_{k+2}^\top u_{k+2}$ will still decrease every two steps, then it returns to the above cases when $|x_k y_k - 1| > R_2$. Then we analyze the case where $u_{k+2}^\top u_{k+2}$ will increase after two steps. For the first n s.t. $x_{k+2n} y_{k+2n} > R_2$, if we have $|r_{k+2n}(R_2)| < |r_k(R_2)|$, this implies $u_{k+2n}^\top u_{k+2n} < u_k^\top u_k$ which can be either absorbed in the above cases, or eventually fall into the following case. For the first n s.t. $x_{k+2n} y_{k+2n} > R_2$, consider $|r_{k+2n}(R_2)| > |r_k(R_2)|$, with $|r_{k+2i}| \ge 1$ for all $i < n$ and $u_{k+2n}^\top u_{k+2n} > u_k^\top u_k$. Then if such process repeats, $|r_{k+2n}(R_2)|$ will be larger and larger until GD either starts to decrease in $|xy - 1| \le R_2$ or enters the two-step decreasing region of $u_k^\top u_k$ in $|xy - 1| \le R_2$ (we can use $|xy - 1| \ge R_1$ to represent this region; however, R_1 is just a bound, meaning the actual region is larger), which will lead to $|r_{k+2n}| < 1$ inside this region due to the same reasoning as the previous paragraph. Detailed characterization of this stage can be seen in the proof of limiting sharpness.

□

D Proofs of Theorem 2 for function (3)

Let $b = 3$ in this section. Let $p(s) = \sum_{i=0}^{b-1} s^i$. Then

$$1 - (xy)^b = (1 - xy) p(xy).$$

Apart from the equations in Appendix B, we will also use the following two equations in our proofs

$$x_{k+1} y_{k+1} - 1 = (x_k y_k - 1) \left(1 - \frac{h}{b} (x_k y_k)^{b-1} p(x_k y_k) u_k^\top u_k + \frac{h^2}{b^2} (x_k y_k - 1)(x_k y_k)^{2b-2} p(x_k y_k)^2 x_k y_k \right),$$

$$u_{k+1}^\top u_{k+1} = u_k^\top u_k - \frac{h}{b} (x_k y_k - 1) \left((x_k y_k)^{b-1} p(x_k y_k) x_k y_k \left(4 - \frac{h}{b} (x_k y_k)^{b-1} p(x_k y_k) u_k^\top u_k \right) + \frac{h}{b} (x_k y_k)^{2b-2} p(x_k y_k)^2 u_k^\top u_k \right).$$

D.1 Proof of convergence

Proof. By Lemma 4 (with a different choice of null set based on the functions), GD can only converge to the point s.t. $x^2 + y^2 \le 2/h$. Also we remove all the points (which is in a null set) s.t. they converge in finite step. Therefore, $r_k \neq 0$ for all $k \ge 0$.

If $0 < r_0 < 1$, we have $x_1 y_1 > 1$ and for any $x_k y_k > 1$,

$$\begin{aligned}
u_{k+1}^\top u_{k+1} = u_k^\top u_k - \frac{h}{b}(x_k y_k - 1)\Bigg(& (x_k y_k)^{b-1} p(x_k y_k) x_k y_k \left(4 - \frac{h}{b}(x_k y_k)^{b-1} p(x_k y_k) u_k^\top u_k\right) \\
& + \frac{h}{b}(x_k y_k)^{2b-2} p(x_k y_k)^2 u_k^\top u_k \Bigg) \\
\le\ & u_k^\top u_k.
\end{aligned}$$

Assume $r_i > 0$ for $i = 0, \cdots, k$. Consider

$$r_{k+1} = \left(1 - h\frac{\ell_{k+1}}{x_{k+1} y_{k+1} - 1}\left(u_{k+1}^\top u_{k+1} - h\ell_{k+1} x_{k+1} y_{k+1}\right)\right)$$

where $\ell_k = \frac{(x_k y_k)^{b-1}((x_k y_k)^b - 1)}{b}$ and therefore $q_k = \frac{\ell_k}{x_k y_k - 1} = \frac{1}{b}(x_k y_k)^{b-1} p(x_k y_k)$. Moreover, $W(s) = \frac{1}{b} s^{b-1} p(s)$ monotonically increases when $s > 1$, which implies $q_{k+1} \le q_k$. Also,

$$\begin{aligned}
0 < u_{k+1}^\top u_{k+1} - h\ell_{k+1} x_{k+1} y_{k+1} &= (1 + h^2 \ell_k^2) u_k^\top u_k - 4h\ell_k x_k y_k - h\ell_{k+1} x_{k+1} y_{k+1} \\
&\le u_k^\top u_k - h\ell_k x_k y_k + h\frac{C}{b}\ell_k x_k y_k - 3h\ell_k x_k y_k - h\ell_{k+1} x_{k+1} y_{k+1} \\
&\le u_k^\top u_k - h\ell_k x_k y_k.
\end{aligned}$$

Therefore, $r_{k+1} \ge r_k > 0$. Moreover, we have

$$\begin{aligned}
r_{k+1} &= 1 - h\frac{\ell_{k+1}}{x_{k+1} y_{k+1} - 1}\left(u_{k+1}^\top u_{k+1} - h\ell_{k+1} x_{k+1} y_{k+1}\right) \\
&\le 1 - h\frac{\ell_{k+1}}{x_{k+1} y_{k+1} - 1}\left(2x_{k+1} y_{k+1} - h\ell_{k+1} x_{k+1} y_{k+1}\right) \\
&\le 1 - h(2 - h\ell_0) < 1
\end{aligned}$$

where the second inequality follows from the value at $x_{k+1} y_{k+1} = 1$. Then $x_k y_k - 1$ exponentially decreases until it converges.

If $r_0 < 0$, from the upper bound of h, we have $0 < r_1 < 1$, $0 < x_1 y_1 < 1$, and $u_1^\top u_1 \le u_0^\top u_0$ (according to the above discussion for $x_0 y_0 > 1$). Then if $0 < x_k y_k < 1$,

$$\begin{aligned}
u_{k+1}^\top u_{k+1} &= (1 + h^2 \ell_k^2) u_k^\top u_k - 4h\ell_k x_k y_k \\
&= u_k^\top u_k + h|\ell_k|(4x_k y_k + h|\ell_k| u_k^\top u_k) \\
&\le u_k^\top u_k + (4 + C u_k^\top u_k / (b u_0^\top u_0)) h|\ell_k| x_k y_k
\end{aligned}$$

where the inequality follows from $x_k y_k < 1$ and $x_0 y_0 > 1$.

Also, when $0 < x_k y_k < 1$,

$$\begin{aligned}
x_{k+1} y_{k+1} &= x_k y_k + (1 - r_k)(1 - x_k y_k) \\
&= x_k y_k - h\ell_k(u_k^\top u_k - h\ell_k x_k y_k) \\
&\ge x_k y_k - h\ell_k x_k y_k (2 - h\ell_k) \\
&= x_k y_k + h|\ell_k| x_k y_k (2 + h|\ell_k|) \\
&\ge x_k y_k + 2h|\ell_k| x_k y_k.
\end{aligned}$$

We claim that when $u_k^\top u_k \le u_0^\top u_0$ and $0 < x_k y_k < 1$, we have $u_n^\top u_n \le u_0^\top u_0 + C_1$ for all $n \ge k$ and for constant $\frac{10}{3} < C_1 \le \frac{7}{2}$. Otherwise, consider N s.t. $u_n^\top u_n \le u_0^\top u_0 + C_1$ for $k \le n \le N$ and

$u_{N+1}^\top u_{N+1} \geq u_0^\top u_0 + C_1$. If $0 < x_n y_n < 1$,

$$\begin{aligned} r_n &= 1 - h\frac{\ell_n}{x_n y_n - 1}(u_n^\top u_n - h\ell_n x_n y_n) \\ &\geq 1 - h(u_n^\top u_n - h\ell_n x_n y_n) \\ &\geq 1 - \frac{C}{4(u_0^\top u_0 + 4)}(u_n^\top u_n - h\ell_n x_n y_n) \\ &\geq 1 - \frac{u_0^\top u_0 + C_1 + h|\ell_n| x_n y_n}{u_0^\top u_0 + 4} \\ &\geq 1 - \frac{u_0^\top u_0 + 15/4}{u_0^\top u_0 + 4} > 0 \end{aligned}$$

where the first inequality follows from $q_n \leq 1$ for $0 < x_n y_n < 1$; the second inequality follows from the initial condition and the requirement of h; the last inequality follows from $h|\ell_n| x_n y_n < \frac{1}{4}$ for $0 < x_n y_n < 1$ (can be easily checked from the initial condition that $h \leq \frac{1}{8}$, and the rest follows from analyzing the expression of the function). Therefore, $0 < x_{n+1} y_{n+1} < 1$ and consequently, $0 < r_n < 1$ for $n = k, \cdots, N$. Especially, consider $n = N$. Then iteratively from the lower bound of $x_{N+1} y_{N+1}$ above, we have

$$\sum_{i=k}^{N} 2h|\ell_i| x_i y_i < 1$$

and thus

$$\sum_{i=k}^{N}(4+C u_i^\top u_i/(b u_0^\top u_0)) h|\ell_i| x_i y_i \leq (4+C(u_0^\top u_0+C_1)/(b u_0^\top u_0)) \sum_{i=k}^{N} h|\ell_i| x_i y_i < \frac{20}{3} \sum_{i=k}^{N} h|\ell_i| x_i y_i < \frac{10}{3} < C_1.$$

Contradiction.

Then, we have $0 < r_n < 1$ for all $n \geq k$. Take $k = 1$, and then we have the monotone decreasing of $1 - x_n y_n$. Also,

$$\begin{aligned} r_n &= 1 - h\frac{\ell_n}{x_n y_n - 1}(u_n^\top u_n - h\ell_n x_n y_n) \\ &\leq 1 - h q_1 (2 x_n y_n - h\ell_n x_n y_n) \\ &\leq 1 - h q_1 x_1 y_1 (2 - h\ell_n) \\ &\leq 1 - h(2-h) q_1 x_1 y_1 < 1. \end{aligned}$$

Thus, GD will converge to $xy = 1$.

□

D.2 Proof of non-EoS

Proof. From the proof of convergence, we know $r_k > 0$ for $k \geq 1$. Thus when $x_k y_k$ is very close to 1, $r_k > 0$. Take the limit and we have

$$\lim_{k \to \infty} r_k = 1 - h u_\infty^\top u_\infty \geq 0.$$

Therefore,

$$S_\infty = u_\infty^\top u_\infty \leq \frac{1}{h}$$

□

D.3 Proof of non-balancing

Proof. Let $h = \frac{C}{(x_0^2+y_0^2+4)(x_0 y_0)^{2b-2}}$. Before the proof, let first consider

$$h\ell_0 = \frac{C}{(x_0^2 + y_0^2 + 4)(x_0 y_0)^{2b-2}} \frac{(x_0 y_0)^{b-1}\left((x_0 y_0)^b - 1\right)}{b} \leq \frac{C x_0 y_0}{b(x_0^2 + y_0^2 + 4)} \leq \frac{C}{2b}.$$

If $r_0 > 0$, from the proof of convergence, we know: $r_k > 0$ for all k, and $\ell_k \geq 0$ for all k. Moreover, we have $x_k y_k \geq x_{k+1} y_{k+1}$, and consequently $\ell_k \leq \ell_0$ for all k (by the monotone decreasing of this function; see details in the proof of convergence). Then

$$\begin{aligned} x_{k+1} y_{k+1} &= x_k y_k - h\ell_k (u_k^\top u_k - h\ell_k x_k y_k) \leq x_k y_k - h\ell_k x_k y_k (2 - h\ell_k) \\ &\leq x_k y_k - h\ell_k x_k y_k (2 - \frac{C}{2b}) \end{aligned}$$

Therefore

$$x_0y_0 - 1 = \sum_{k=0}^{\infty} x_ky_k - x_{k+1}y_{k+1} \geq \sum_{k=0}^{\infty} h\ell_kx_ky_k(2 - \frac{C}{2b}).$$

Also, we have

$$\begin{aligned} u_{k+1}^{\top}u_{k+1} &= (1 + h^2\ell_k^2)u_k^{\top}u_k - 4h\ell_kx_ky_k \\ &\geq u_k^{\top}u_k - 4h\ell_kx_ky_k \end{aligned}$$

Then

$$u_0^{\top}u_0 - u_{\infty}^{\top}u_{\infty} = \sum_{k=0}^{\infty} u_k^{\top}u_k - u_{k+1}^{\top}u_{k+1} \leq \sum_{k=0}^{\infty} 4h\ell_kx_ky_k \leq \frac{8b}{4b - C}(x_0y_0 - 1) \leq \frac{2b}{b-1}(x_0y_0 - 1)$$

Then

$$\begin{aligned} (x_{\infty} - y_{\infty})^2 = u_{\infty}^{\top}u_{\infty} - 2x_{\infty}y_{\infty} &\geq u_0^{\top}u_0 - 2x_{\infty}y_{\infty} - \frac{8b}{4b - C}(x_0y_0 - 1) \\ &= u_0^{\top}u_0 - 2x_0y_0 + 2x_0y_0 - 2 - \frac{8b}{4b - C}(x_0y_0 - 1) \\ &= (x_0 - y_0)^2 - \frac{2C}{4b - C}(x_0y_0 - 1). \end{aligned}$$

If $r_0 < 0$, from the proof of convergence, we have: $r_k > 0$ for $k \geq 1$, and $u_{k+1}^{\top}u_{k+1} \geq u_k^{\top}u_k$ for $k \geq 1$. Thus

$$u_0^{\top}u_0 - u_{\infty}^{\top}u_{\infty} \leq u_0^{\top}u_0 - u_1^{\top}u_1 \leq 4h\ell_0x_0y_0 \leq \frac{2C}{b}x_0y_0$$

Then

$$\begin{aligned} (x_{\infty} - y_{\infty})^2 = u_{\infty}^{\top}u_{\infty} - 2x_{\infty}y_{\infty} &\geq u_0^{\top}u_0 - 2x_{\infty}y_{\infty} - \frac{2C}{b}x_0y_0 \\ &= u_0^{\top}u_0 - 2x_0y_0 + 2x_0y_0 - 2 - \frac{2C}{b}x_0y_0 \\ &= (x_0 - y_0)^2 + 2(x_0y_0 - 1) - \frac{2C}{b}x_0y_0. \end{aligned}$$

□

D.4 Supplementary lemmas

Lemma 9. *Under the assumption of Theorem 2, $M_3 > 3$.*

Proof. By the assumption, we have

$$\begin{aligned} &\left(1 + \left(\frac{C}{(x_0^2 + y_0^2 + 4)(x_0y_0)^{2b-2}}\right)^2 \ell_0^2\right)x_0y_0 - \frac{C}{(x_0^2 + y_0^2 + 4)(x_0y_0)^{2b-2}}\ell_0(x_0^2 + y_0^2) \\ &= \left(1 + \left(\frac{C\left((x_0y_0)^b - 1\right)}{b(x_0^2 + y_0^2 + 4)(x_0y_0)^{b-1}}\right)^2\right)x_0y_0 - \frac{C\left((x_0y_0)^b - 1\right)}{b(x_0^2 + y_0^2 + 4)(x_0y_0)^{b-1}}(x_0^2 + y_0^2) \\ &\geq x_0y_0 - \frac{C(x_0^2 + y_0^2)\left((x_0y_0)^b - 1\right)}{b(x_0^2 + y_0^2 + 4)(x_0y_0)^{b-1}} \end{aligned}$$

Since $b \geq 3$, when $C = 3$, we have

$$x_0y_0 - \frac{C(x_0^2 + y_0^2)\left((x_0y_0)^b - 1\right)}{b(x_0^2 + y_0^2 + 4)(x_0y_0)^{b-1}} > 0.$$

□

Lemma 10 (stepsize)**.** *When $x_0^2 + y_0^2 \geq 4\left(\sqrt{2} + 2\right)C_1$, the learning rate bound in Theorem 2*

$$\frac{2}{(u_0^{\top}u_0 + C_1)(x_0y_0)^{2b-2}} \leq h \leq \frac{3}{(u_0^{\top}u_0 + C_1)(x_0y_0)^{2b-2}}$$

satisfies

$$\frac{2}{S_0} < \frac{2}{(u_0^{\top}u_0 + C_1)(x_0y_0)^{2b-2}}, \text{ and } \frac{4}{S_0} \leq \frac{3}{(u_0^{\top}u_0 + C_1)(x_0y_0)^{2b-2}}.$$

Proof. Consider the Hessian $\nabla^2 f(x,y)$. The trace is

$$\begin{aligned}
\operatorname{tr}\nabla^2 f(x,y) &= (1+(1-1/b)(1-1/(xy)^b))(x^2+y^2)(xy)^{2b-2} \\
&> (1+(1-1/b)(1-\frac{1}{4^{\frac{b}{2b-2}}}))(x^2+y^2)(xy)^{2b-2} \\
&\geq \left(1+\frac{2}{3}\left(1-\frac{1}{2\sqrt{2}}\right)\right)(x^2+y^2)(xy)^{2b-2} \\
&> \frac{4}{3}(x^2+y^2)(xy)^{2b-2}.
\end{aligned}$$

The determinant is

$$\det\nabla^2 f(x,y) = -\frac{(xy)^{2b}\left((xy)^b-1\right)\left(-(xy)^b+b\left(4(xy)^b-2\right)+1\right)}{b^2x^2y^2}.$$

When $xy > 1$, $\det\nabla^2 f(x,y) < 0$. Therefore, initial sharpness

$$S_0 > \operatorname{tr}\nabla^2 f(x_0,y_0).$$

Then when $x_0^2+y_0^2 \geq 4\left(\sqrt{2}+2\right)C_1 = 16\left(\sqrt{2}+2\right)$, we have

$$S_0 > \left(1+\frac{2}{3}\left(1-\frac{1}{2\sqrt{2}}\right)\right)(x_0^2+y_0^2)(x_0y_0)^{2b-2} \geq \frac{4}{3}(x_0^2+y_0^2+C_1)(x_0y_0)^{2b-2},$$

and thus the lower bound of h is greater than $\frac{2}{S_0}$, actually $\frac{8}{3S_0}$, and the upper bound of h is greater than $\frac{4}{S_0}$. □

E Experiments

For the experimental setup, we test on CIFAR-10 and MNIST, and consider neural network model with one hidden layer of width $N_1 = 200$ and no bias. The input dimension N_0 is $32 \times 32 \times 3 = 3072$ for CIFAR-10-1k, and $28 \times 28 = 784$ for MNIST. The output dimension N_2 is 10. Therefore the weight matrices for each layer are $W_1 \in \mathbb{R}^{N_0 \times N_1}$, $W_2 \in \mathbb{R}^{N_1 \times N_2}$. There are 1000 training data points in our model randomly chosen in CIFAR-10 and MNIST.

For the training, we use full batch gradient descent without weight decay and momentum. The weight initialization follows the default uniform distribution on interval $[-\frac{1}{\sqrt{N_{i-1}}}, \frac{1}{\sqrt{N_{i-1}}}]$ in PyTorch for the ith layer, with a rescaling of the two weights such that $\|W_1\|_F = 6$ and $\|W_2\|_F = 20$, which falls into the sharp/unbalanced initialization.