Large Catapults in Momentum Gradient Descent with Warmup

An Empirical Study

NeurIPS 2023 M3L Workshop Oral

Prin Phunyaphibarn* (KAIST Math), **Junghyun Lee*** (KAIST AI),
Bohan Wang (USTC), Huishuai Zhang (Microsoft Research Asia), Chulhee "Charlie" Yun (KAIST AI)

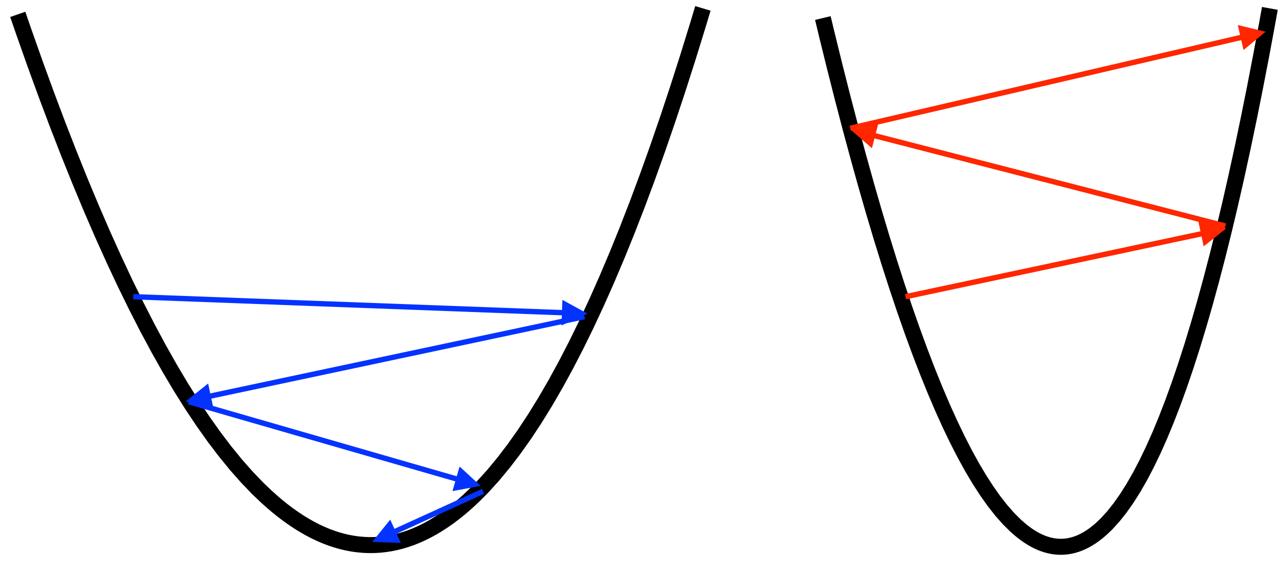
Preliminaries

Optimization with Large Learning Rates

- **Maximum stable sharpness (MSS).** For a quadratic loss, the threshold at which the optimization algorithm diverges if its sharpness goes above it

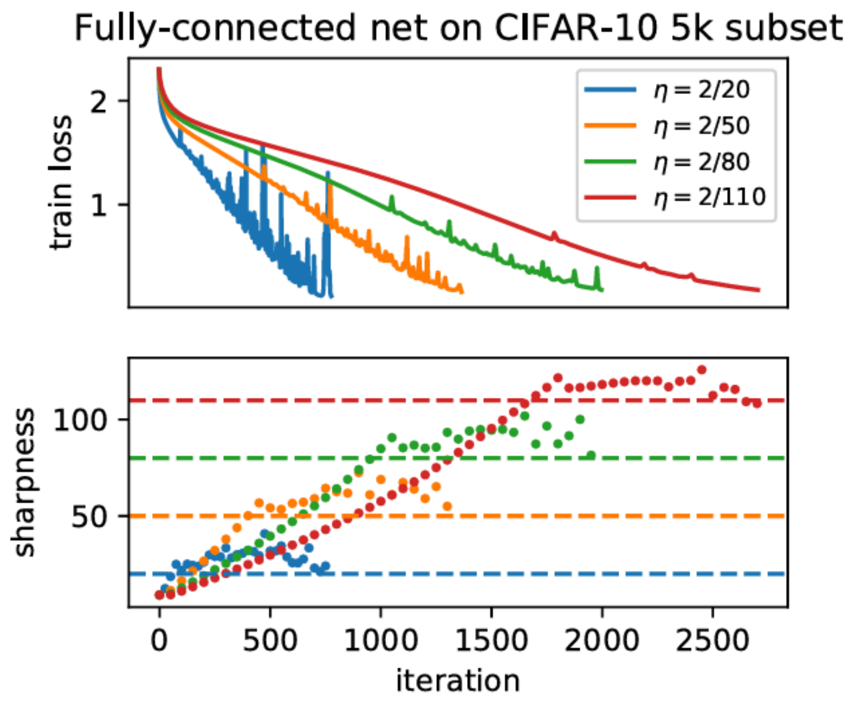
  [Cohen et al., ICLR'21] The MSS of GD with momentum at time t is $\dfrac{2(1+\beta)}{\eta_t}$

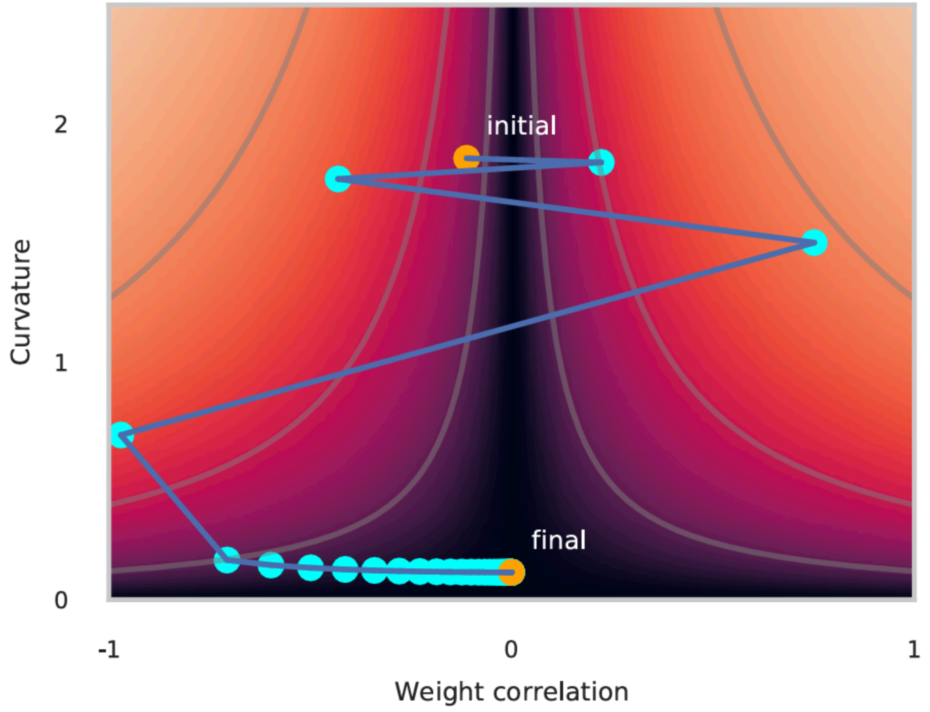
- Interesting, *non-monotone* behaviors:
  - Edge of Stability (EoS) [Cohen et al., ICLR'21]
  - Catapults [Lewkowycz et al., 2020]
  - Balancing Effect [Wang et al., ICLR'22]



- Implicit bias of moderate/large learning rates:

  [Li et al., NeurIPS'19; Wu et al., ICLR'21; Damian et al., NeurIPS'21]





Preliminaries

Learning rate warmup

- To stably train with high learning rate η_f, we use ***learning rate warmup***
- The use of warmup (and its effectiveness) has been studied extensively [Gotmare et al., ICLR'19; Liu et al., ICLR'20]
- **Linear warmup.** Starting from an small initial learning rate η_i, linearly increase the learning rate to η_f over the prescribed warmup period T_{warmup}:

$$\eta_t = \eta_i + \frac{\eta_f - \eta_i}{T_{warmup}} t$$

- During the warmup, we have a *decaying MSS curve*!

Linear Diagonal Networks

Motivating example

- ***Linear diagonal network (LDN):***

$$f(x; u, v) := < \overset{\triangleq w}{\boxed{u \odot u - v \odot v}}, x > , \quad x, u, v \in \mathbb{R}^d$$

- ***Sparse regression:*** the ground truth $w^\star$ is assumed to be sparse!
  - Training samples. $x_n \sim \mathcal{N}_d(\mu, \sigma^2 I), \ y_n = \langle w^\star, x_n \rangle$
- ***Initialization:*** $u_0 = v_0 = \alpha \cdot 1$, with $\alpha > 0$ being the *initialization scale*

We want an implicit bias towards sparse $w_T = u_T \odot u_T - v_T \odot v_T$

Linear Diagonal Networks

Known results

- Gradient flow [Woodworth et al., COLT'20; Pesme & Flammarion, NeurIPS'23]

: minimum ℓ_1-norm solution as $\alpha \to 0$, and minimum ℓ_2-norm solution as $\alpha \to \infty$

: saddle-hopping dynamics

- Stochastic gradient flow [Pesme et al., NeurIPS'21]

: stochasticity better generalization capability then gradient flow

- (S)GD with finite learning rate [Nacson et al., ICML'22; Even et al., NeurIPS'23]

: finite learning rate gives better generalization capability than flow regime, even at large α

What happens if we add momentum???

Linear Diagonal Networks

Known results

- Gradient flow [Woodworth et al., COLT'20; Pesme & Flammarion, NeurIPS'23]

: minimum ℓ_1-norm solution as $\alpha \to 0$, and minimum ℓ_2-norm solution as $\alpha \to \infty$

: saddle-hopping dynamics

- Stochastic gradient flow [Pesme et al., NeurIPS'21]

: stochasticity better generalization capability then gradient flow

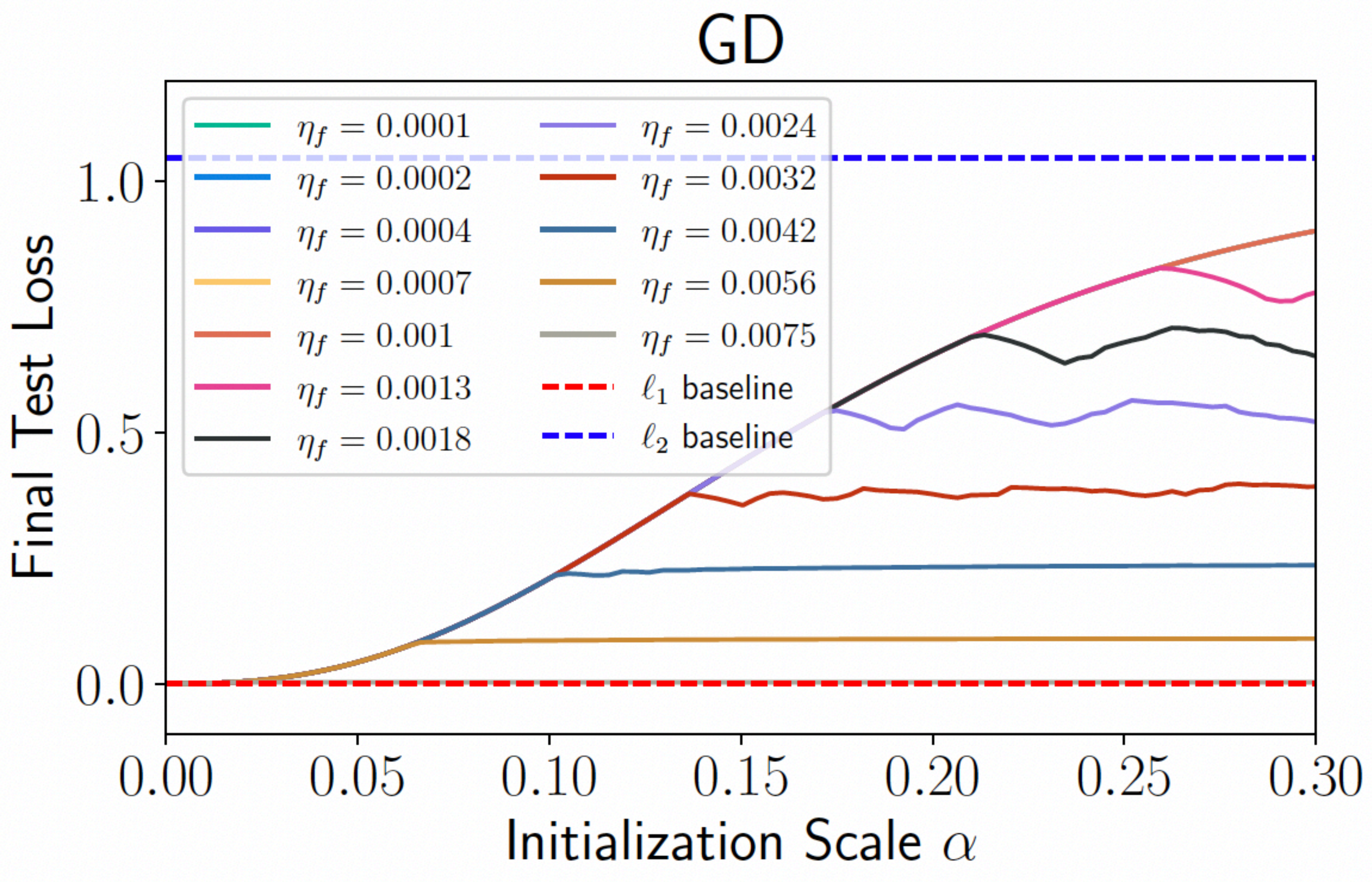
- (S)GD with finite learning rate [Nacson et al., ICML'22; Even et al., NeurIPS'23]

: finite learning rate gives better generalization capability than flow regime, even at large α

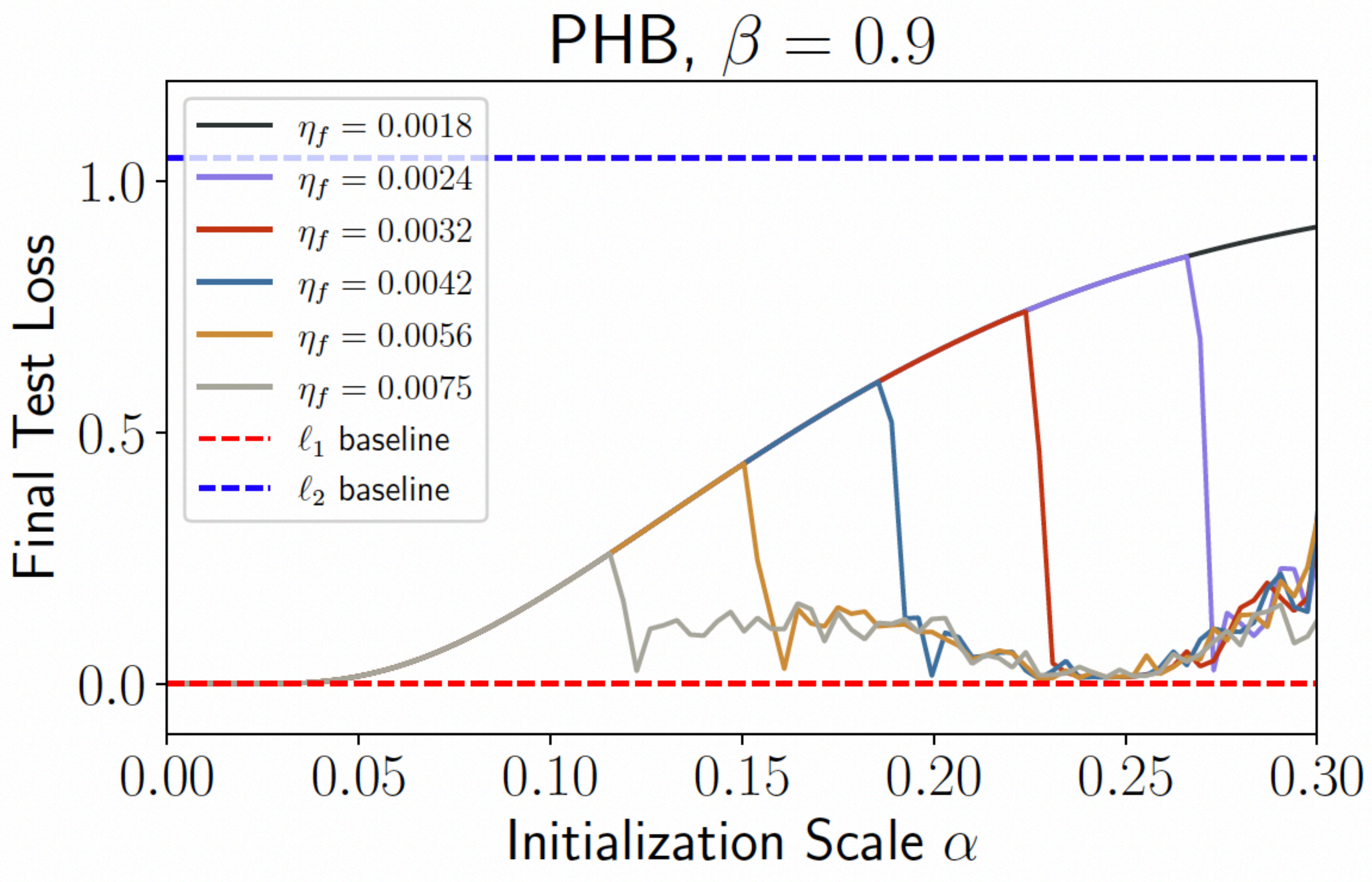
What happens if we add momentum???

Linear Diagonal Networks

Momentum induces fundamentally different implicit bias!



(a) GD



(b) PHB

[Nacson et al., ICML'22]

Our new observation!!

Linear Diagonal Networks

Closer look reveals catapults!

Catapult Mechanism

Definition

- **Catapult.** “a sharp increase in loss, followed by a decrease that forms a single spike in the training loss, coupled with a rapid sharpness reduction”
- So far studied only for GD [Lewkowycz et al., 2020; Meltzer & Liu, 2023; Zhu et al., 2022; 2023]

Some properties:

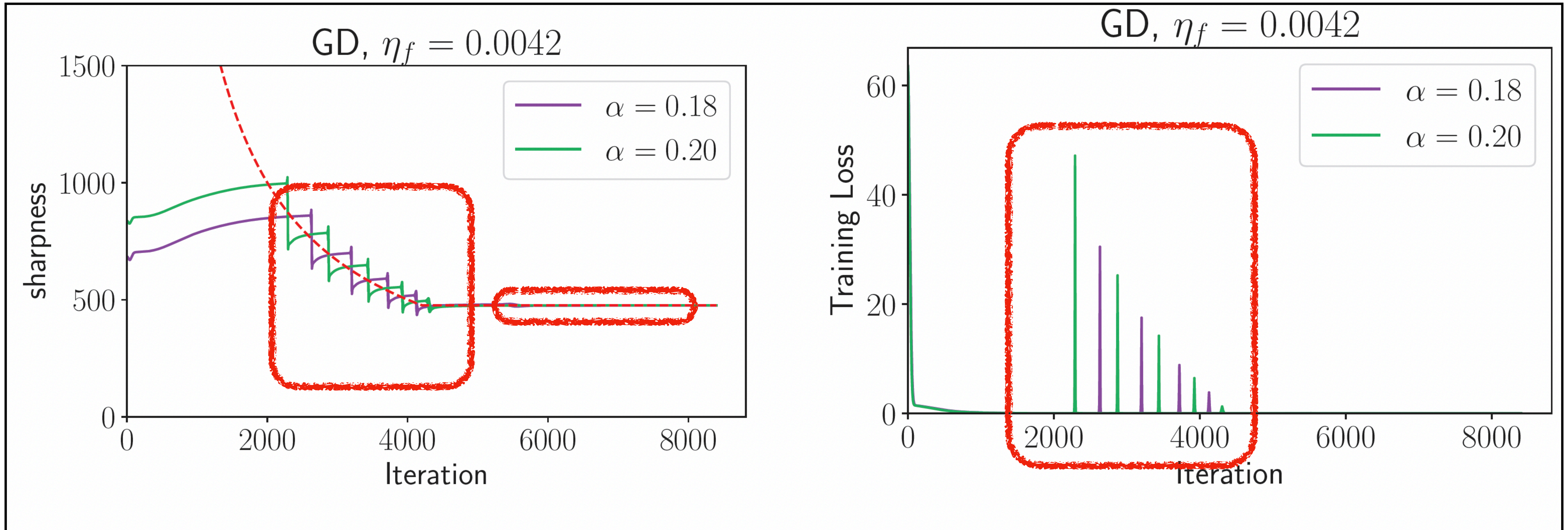
- Iterates are “catapulted” to flatter minima!
- The catapult lasts for a very short time.

Our LDN (sharpness/loss) plots resemble catapult(s)!

Catapult Mechanism for LDNs

Gradient descent

- Sharpness closely follows the MSS curve with **multiple, *small* catapults**
- The final converged sharpness is *just below* the MSS of the final learning rate

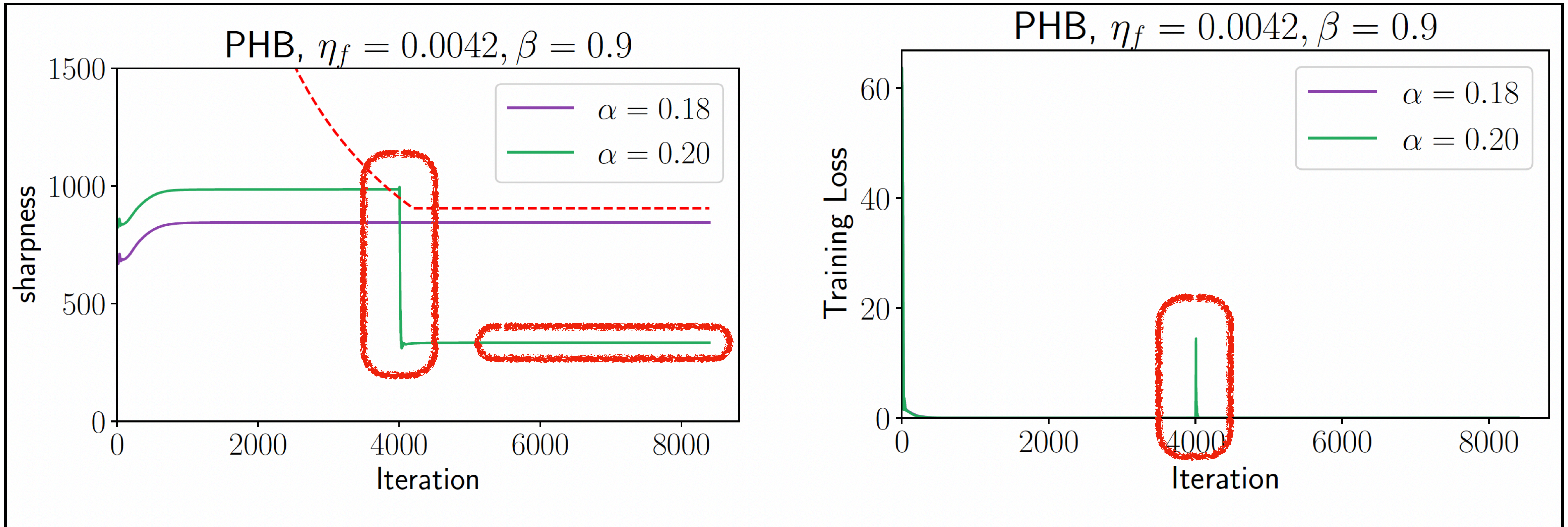


(c) GD Sharpness (d) GD Train Loss

Catapult Mechanism for LDNs

Gradient descent *with momentum*

- Sharpness stays constant until it crosses the MSS curve, at which point there is a **single, large catapult**
- The final converged sharpness is *well below* the MSS of the final learning rate



(e) PHB Sharpness (f) PHB Train Loss

Towards the “Why” of Large Catapults

Toy example

- Consider a 2D toy loss function: $f(x,y) = \dfrac{x^2}{2y}, \quad y > 0.$
- Intuitions behind this toy example:
  - x-direction: *unstable* direction
  - y-direction: *sharpness* changing direction
- “Analogue” of the *self-stabilization* mechanism [Damian et al., ICLR’23]
- **Future work.** Considering more “realistic” toy losses (e.g., quadratic regression?)

Towards the "Why" of Large Catapults

Toy example!

- The trajectory resembles the *self-stabilization* [Damian et al., ICLR'23] for GD:

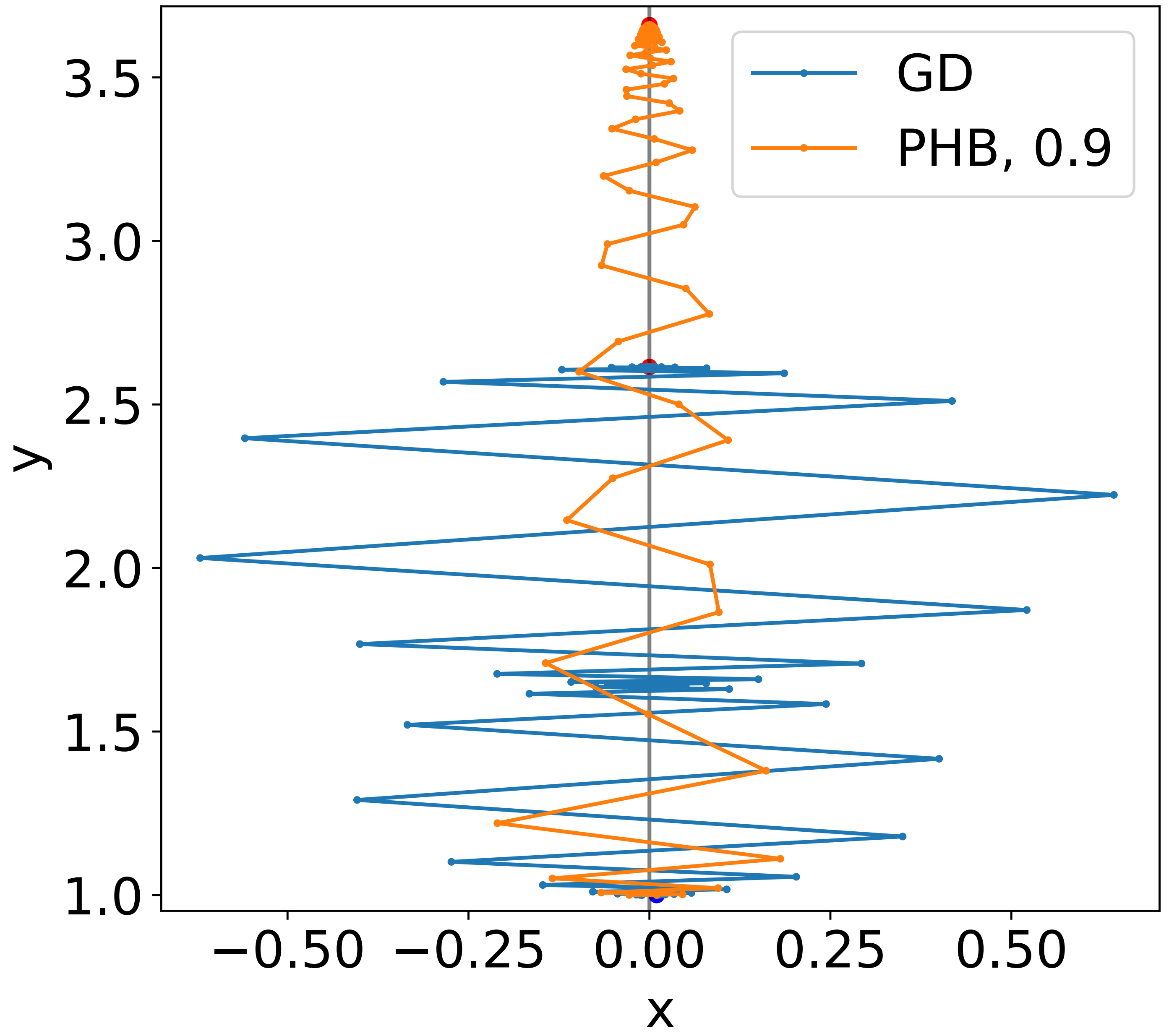
  1. **Progressive Sharpening**[1]
  2. **Blowup**

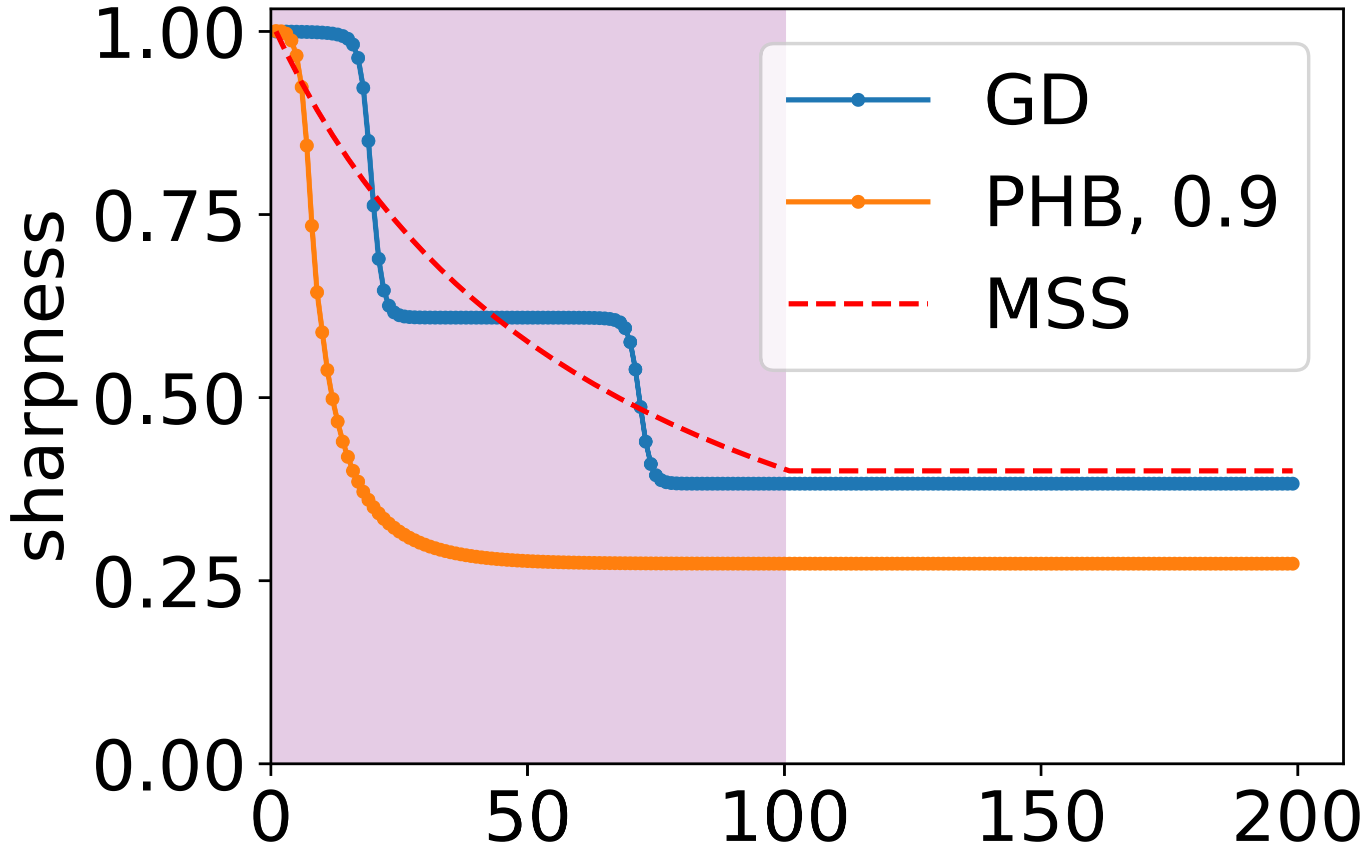
     When the sharpness > MSS, a (locally) divergent dynamics in the x direction causes a sharp increase in the loss, *while* shooting the iterates in $+y$ direction
  3. ***Self-Stabilization***

     Movement in $+y$ direction *stabilizes* the dynamics in the x direction, *and* decreases sharpness
  4. **Return to Stability**

     When the sharpness drops below MSS, the iterates converges locally

- A single catapult is basically step 2~3!





[1]This may not occur depending on the problem setting, such as model complexity and initialization.

Towards the "Why" of Large Catapults

Toy example!

- The trajectory resembles the *self-stabilization* [Damian et al., ICLR'23] for GD:

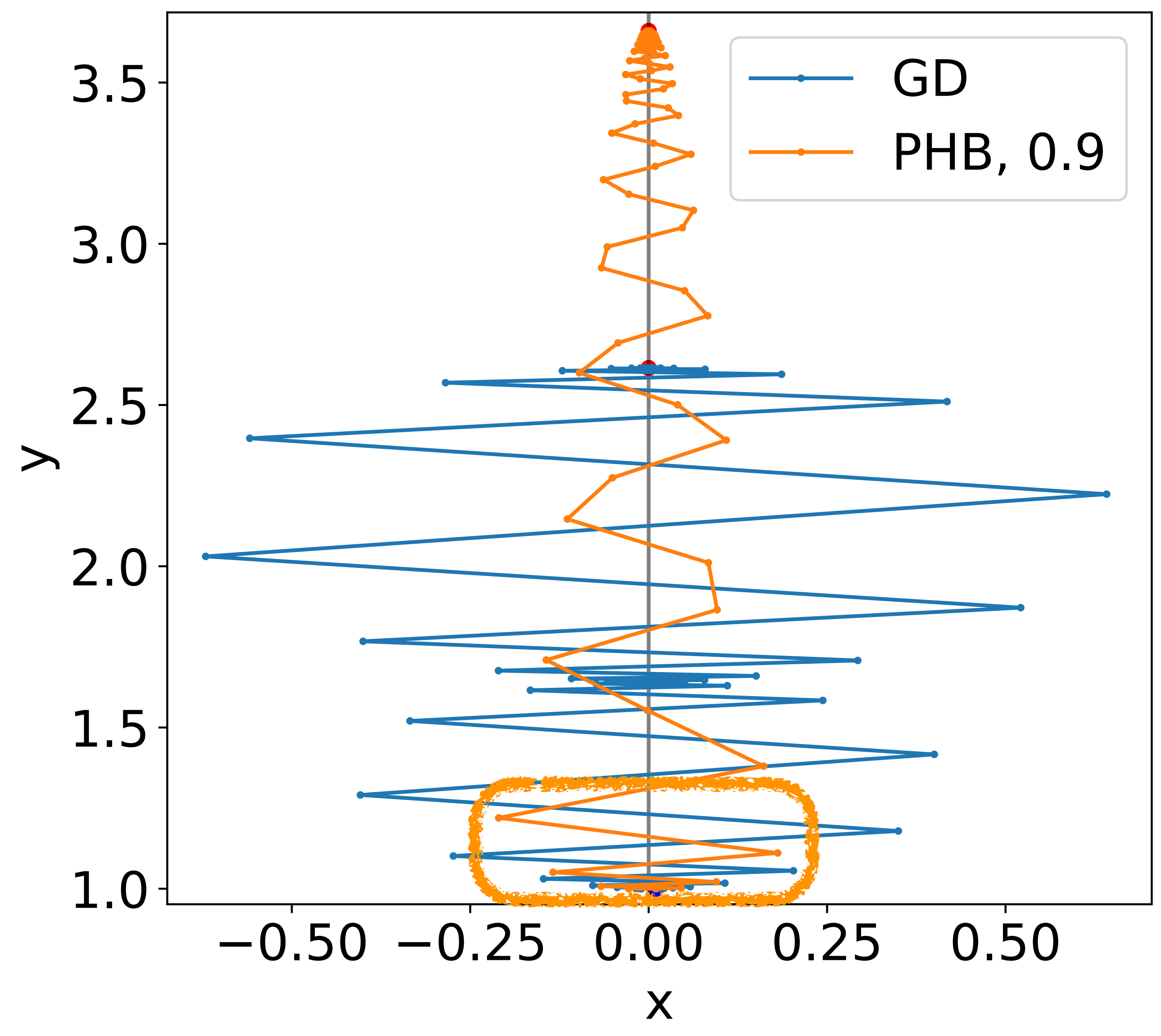
1. **Progressive Sharpening[1]**
2. **Blowup**

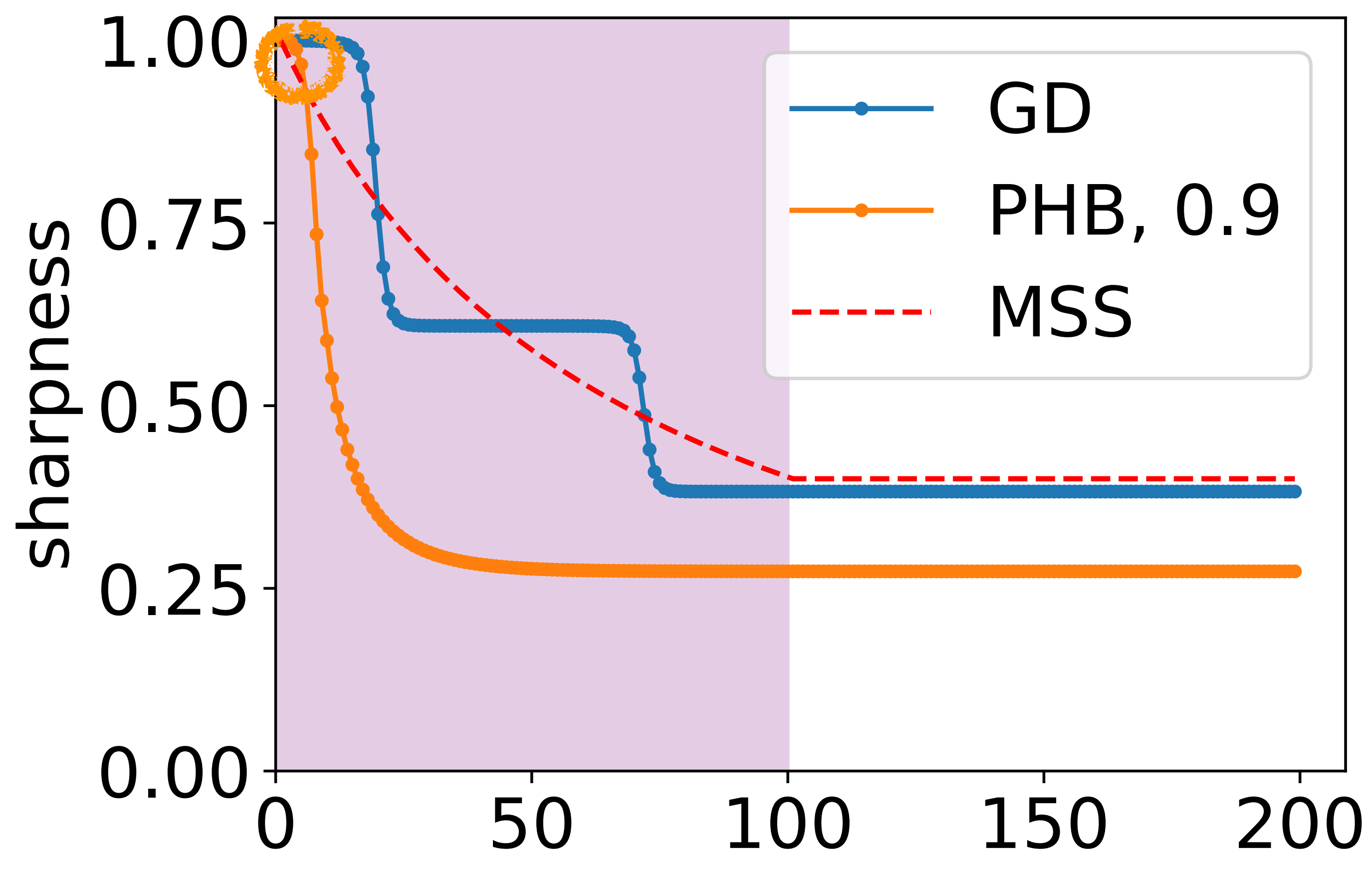
   When the sharpness > MSS, a (locally) divergent dynamics in the x direction causes a sharp increase in the loss, *while* shooting the iterates in $+y$ direction
3. ***Self-Stabilization***

   Movement in $+y$ direction *stabilizes* the dynamics in the x direction, *and* decreases sharpness
4. **Return to Stability**

   When the sharpness drops below MSS, the iterates converges locally

- A single catapult is basically step 2~3!





[1]This may not occur depending on the problem setting, such as model complexity and initialization.

Towards the "Why" of Large Catapults

Toy example!

- The trajectory resembles the *self-stabilization* [Damian et al., ICLR'23] for GD:

  1. **Progressive Sharpening**[1]

  2. **Blowup**

     When the sharpness > MSS, a (locally) divergent dynamics in the x direction causes a sharp increase in the loss, *while* shooting the iterates in $+y$ direction

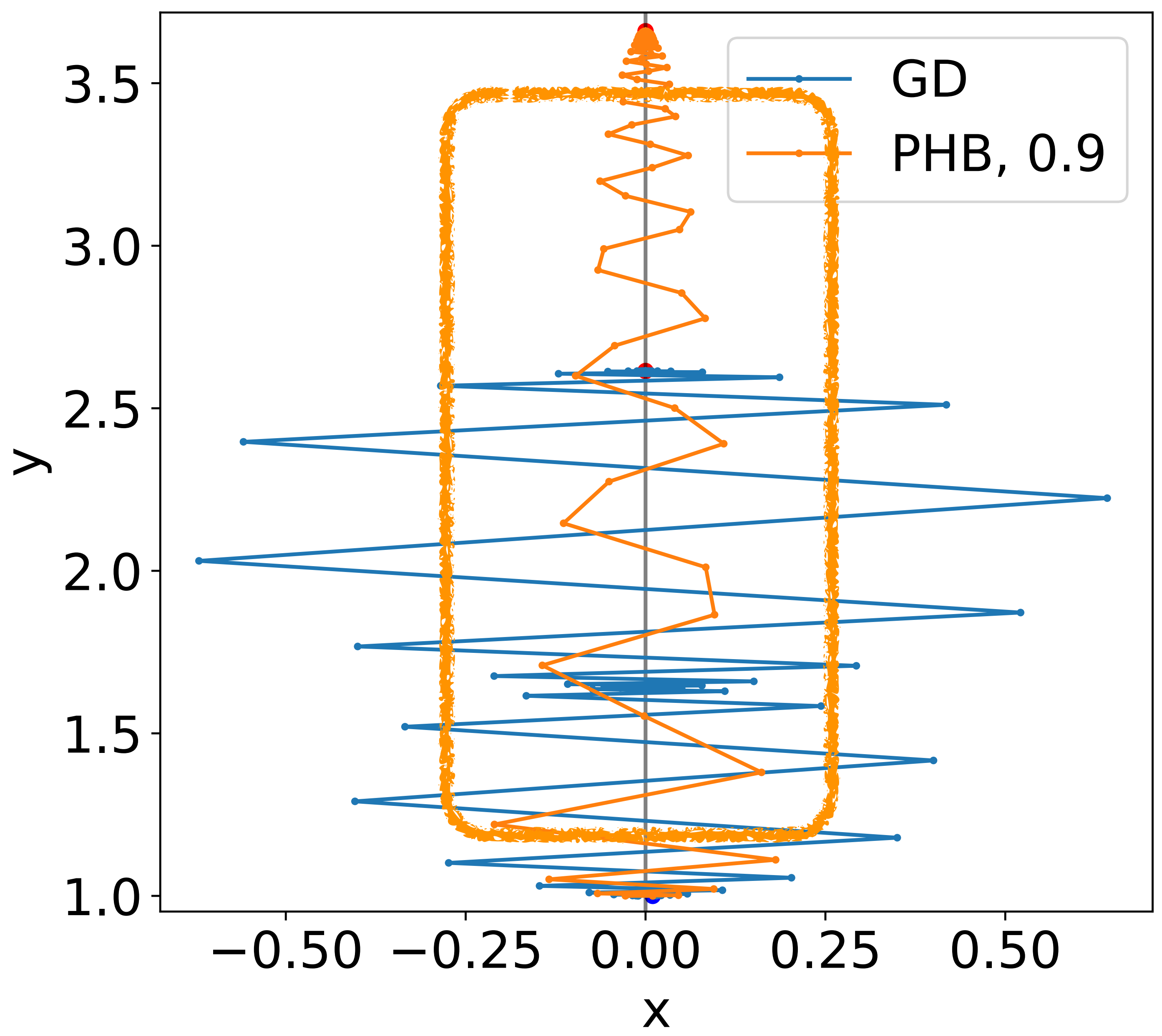
  3. ***Self-Stabilization***

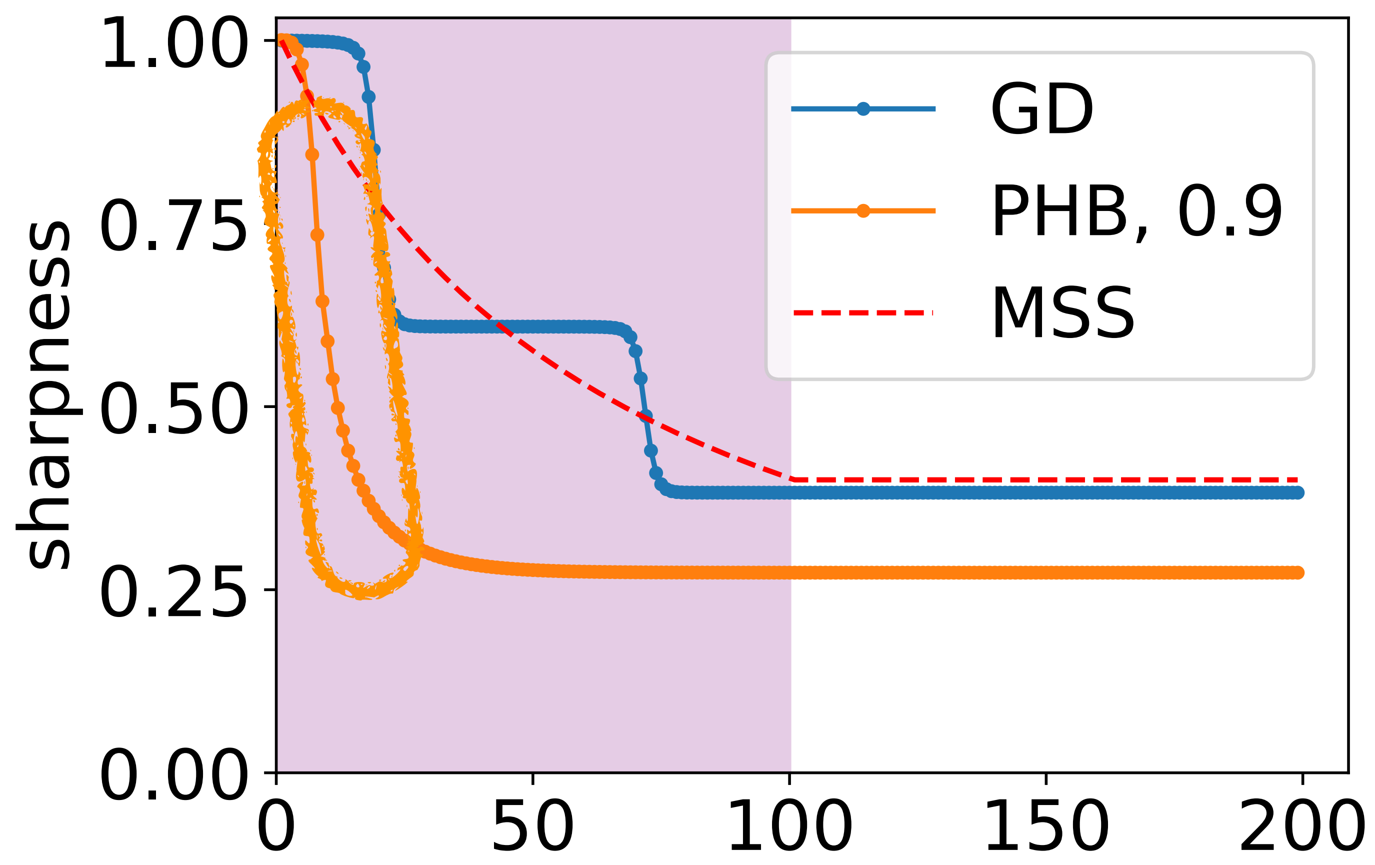
     Movement in $+y$ direction *stabilizes* the dynamics in the x direction, *and* decreases sharpness

  4. **Return to Stability**

     When the sharpness drops below MSS, the iterates converges locally

- A single catapult is basically step 2~3!





[1]This may not occur depending on the problem setting, such as model complexity and initialization.

Towards the “Why” of Large Catapults

Toy example!

- The trajectory resembles the *self-stabilization* [Damian et al., ICLR’23] for GD:

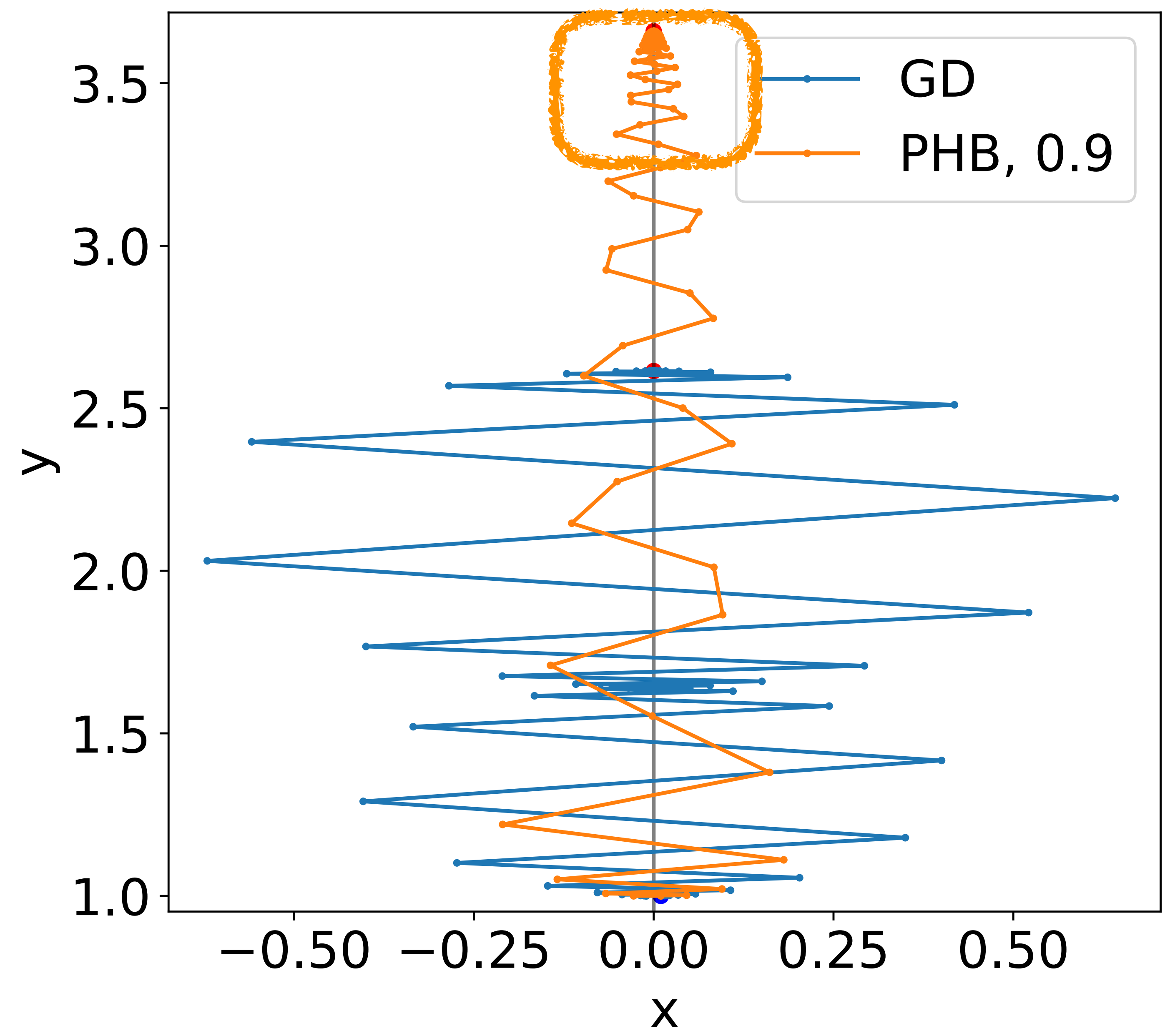
  1. **Progressive Sharpening[1]**
  2. **Blowup**

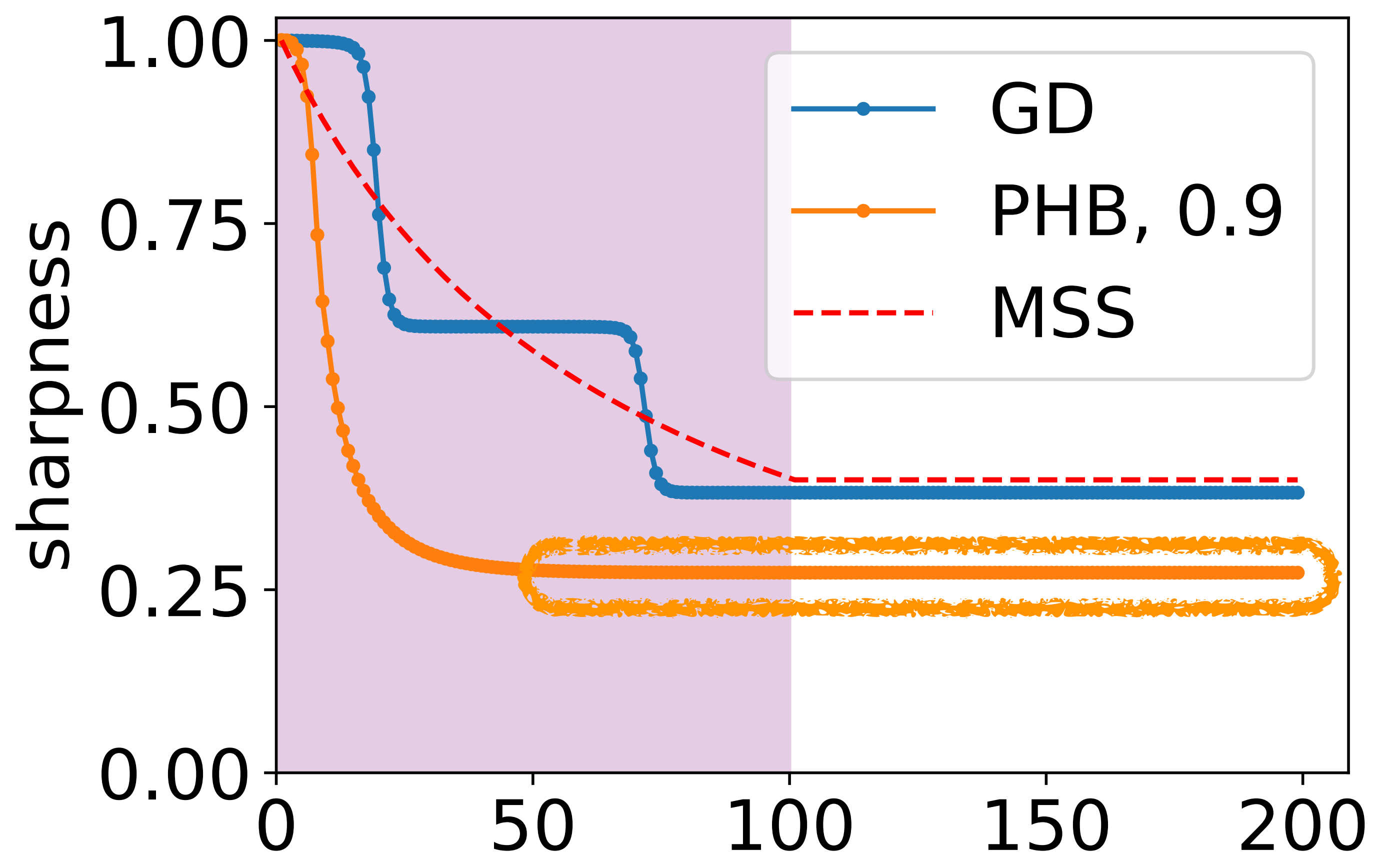
     When the sharpness > MSS, a (locally) divergent dynamics in the x direction causes a sharp increase in the loss, *while* shooting the iterates in $+y$ direction
  3. ***Self-Stabilization***

     Movement in $+y$ direction *stabilizes* the dynamics in the x direction, *and* decreases sharpness
  4. **Return to Stability**

     When the sharpness drops below MSS, the iterates converges locally

- A single catapult is basically step 2~3!



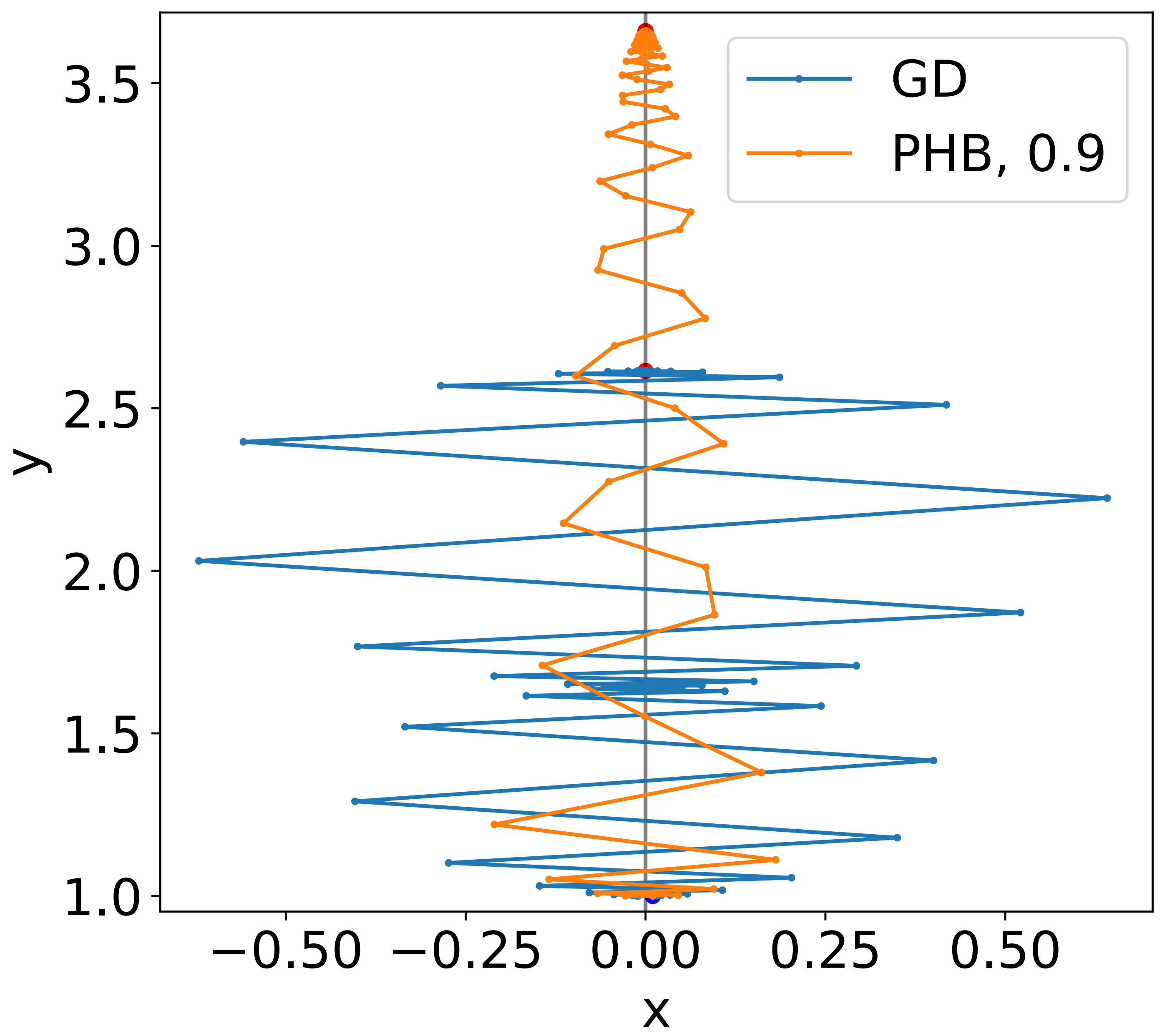


[1]This may not occur depending on the problem setting, such as model complexity and initialization.

Towards the "Why" of Large Catapults

Main Hypotheses

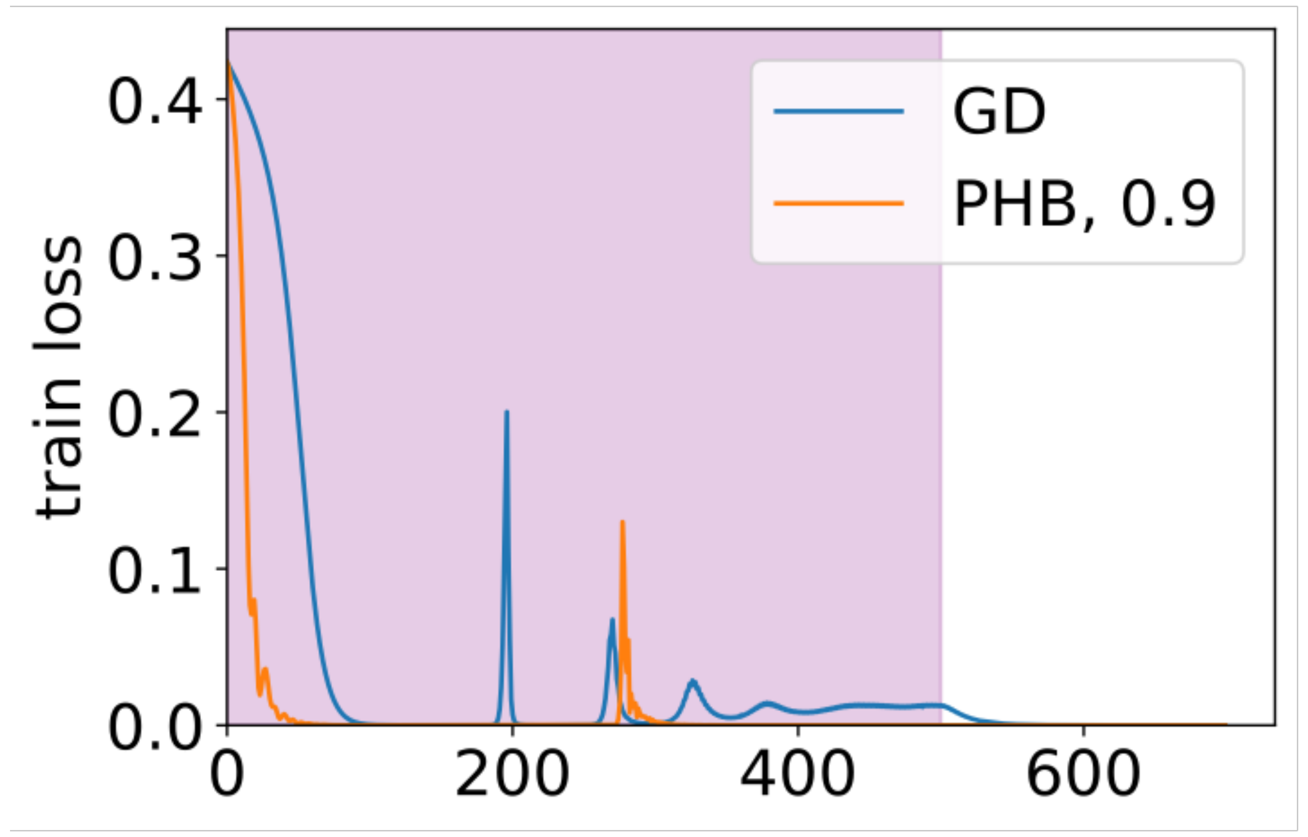
- Momentum *controls* the blow-up from being too large via **dampening effect**
- Momentum *prolongs* the self-stabilization via acceleration in the **direction of the negative gradient of the sharpness**

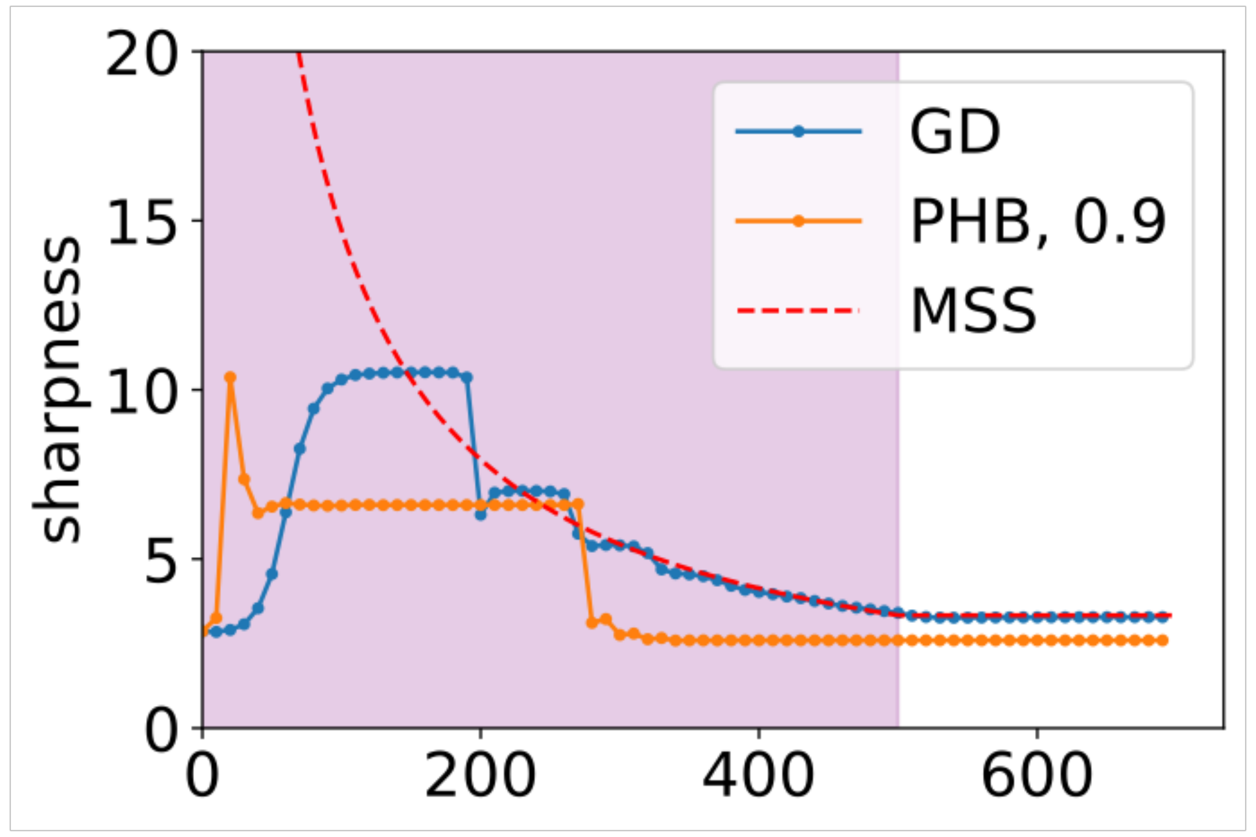


Nonlinear Neural Networks

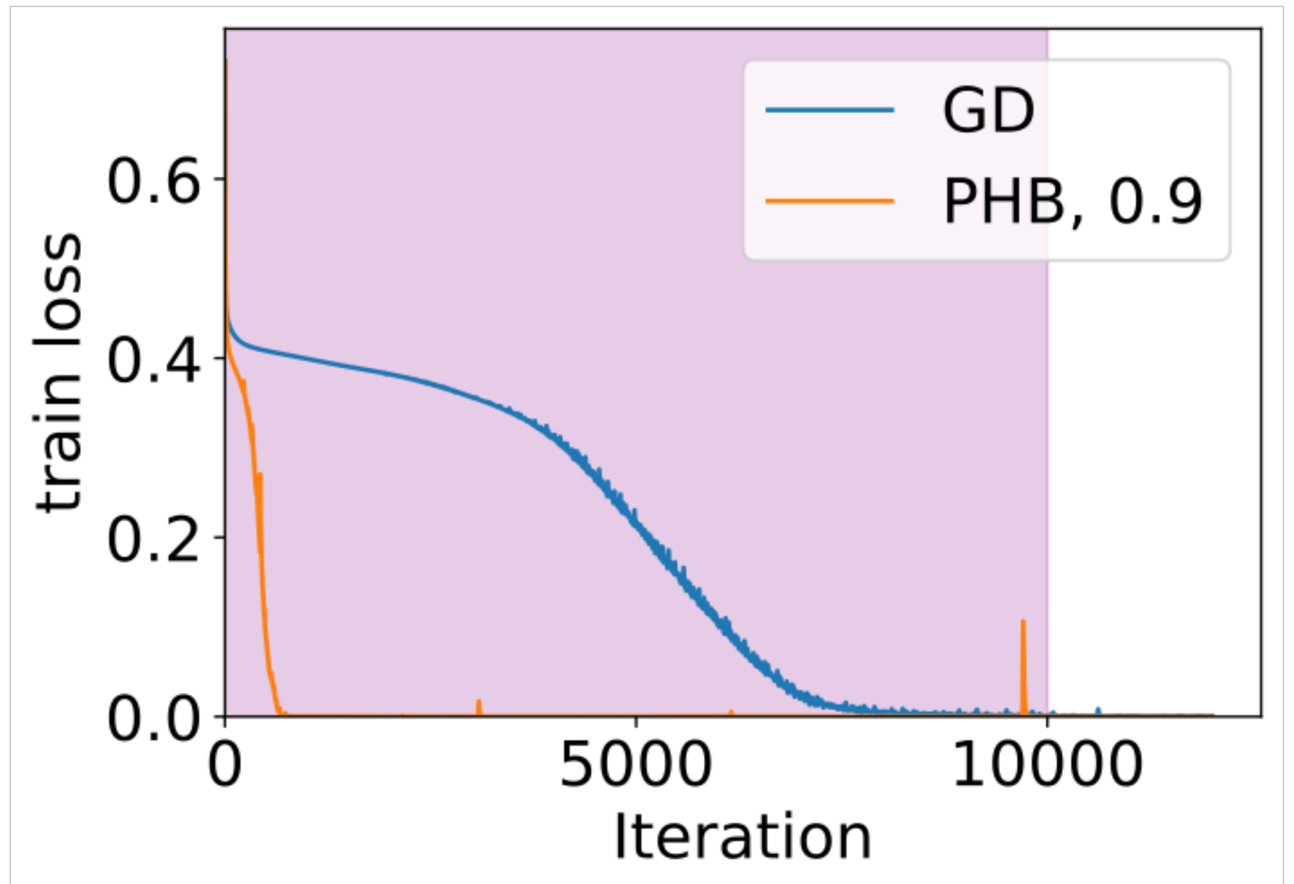
The observations hold for more complex scenarios!

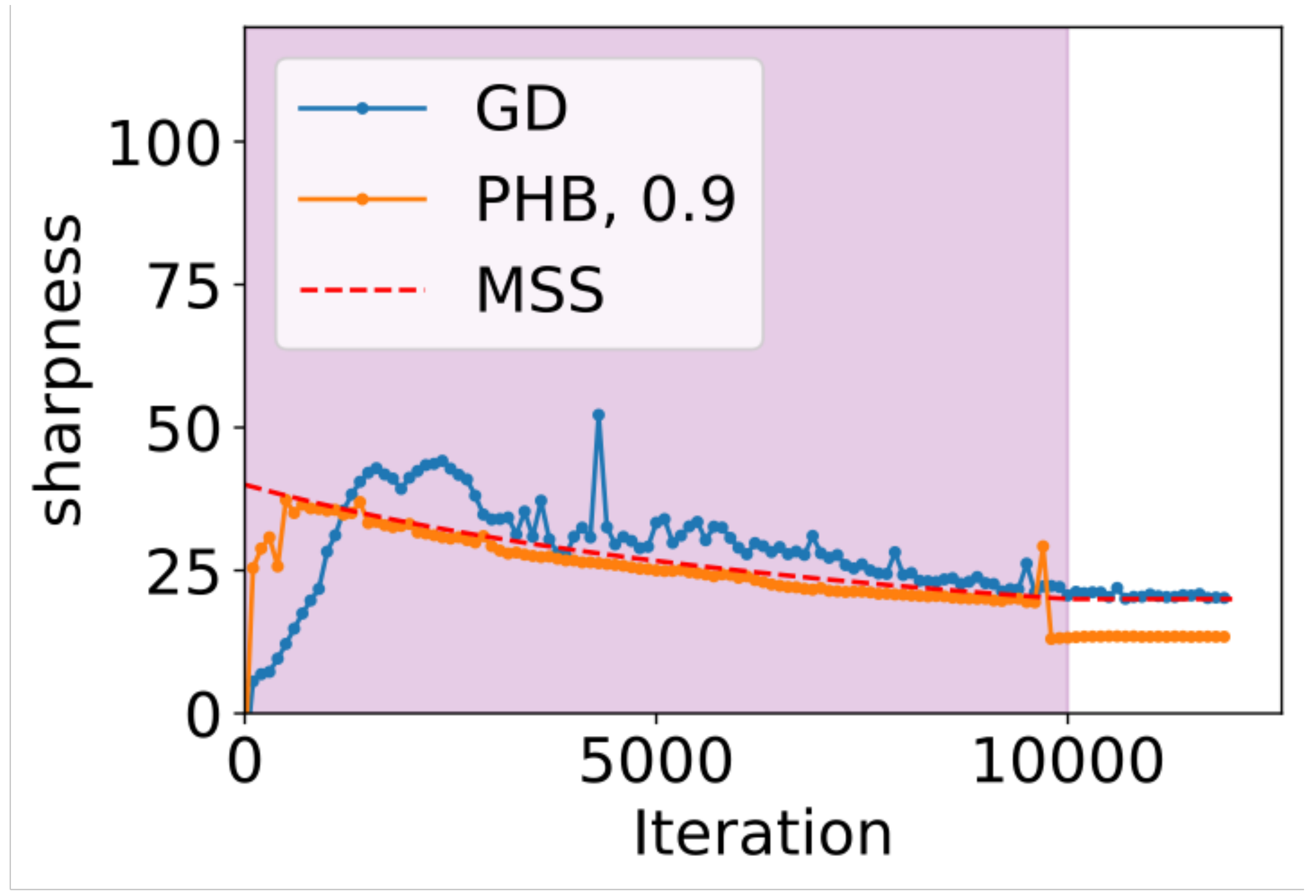
FCN trained on rank-2 (synthetic) dataset:





ResNet20 trained on 1k subset of CIFAR10:

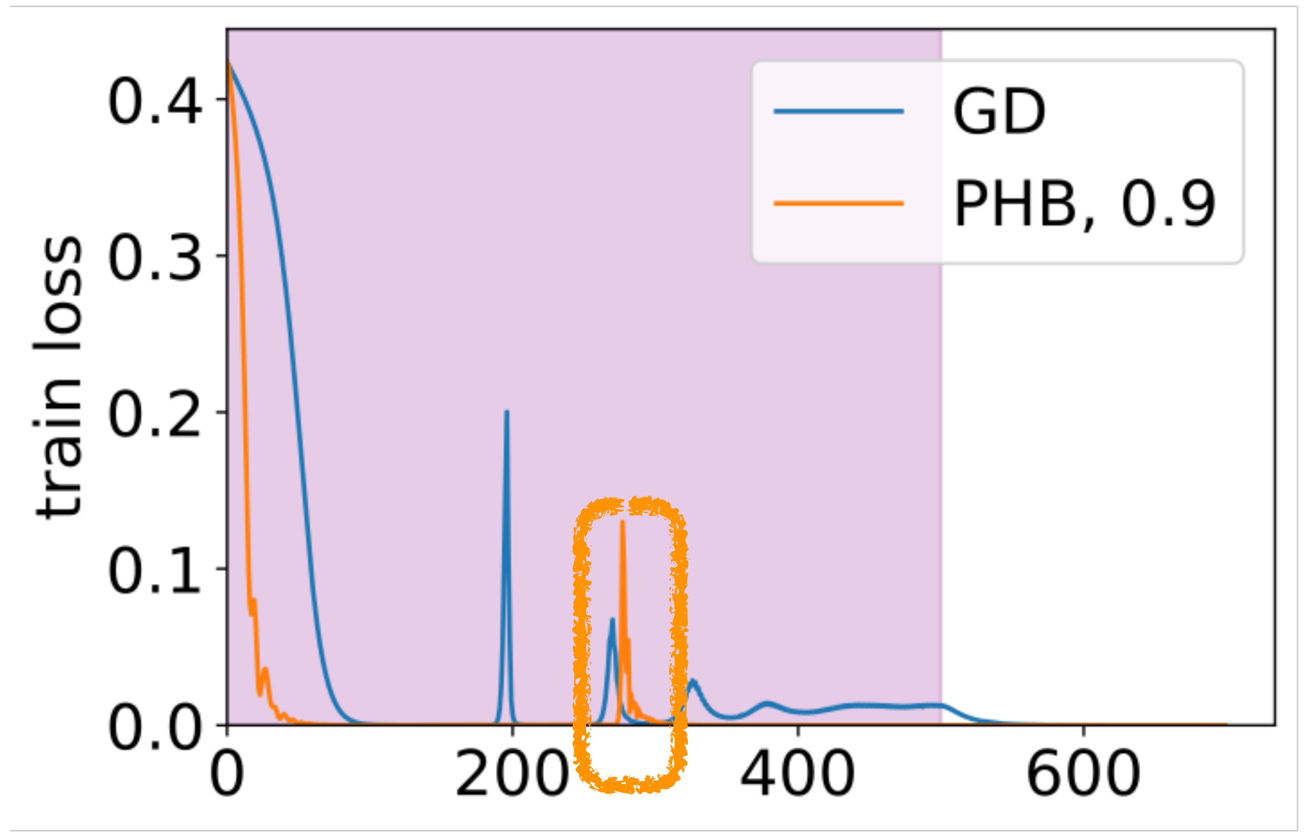


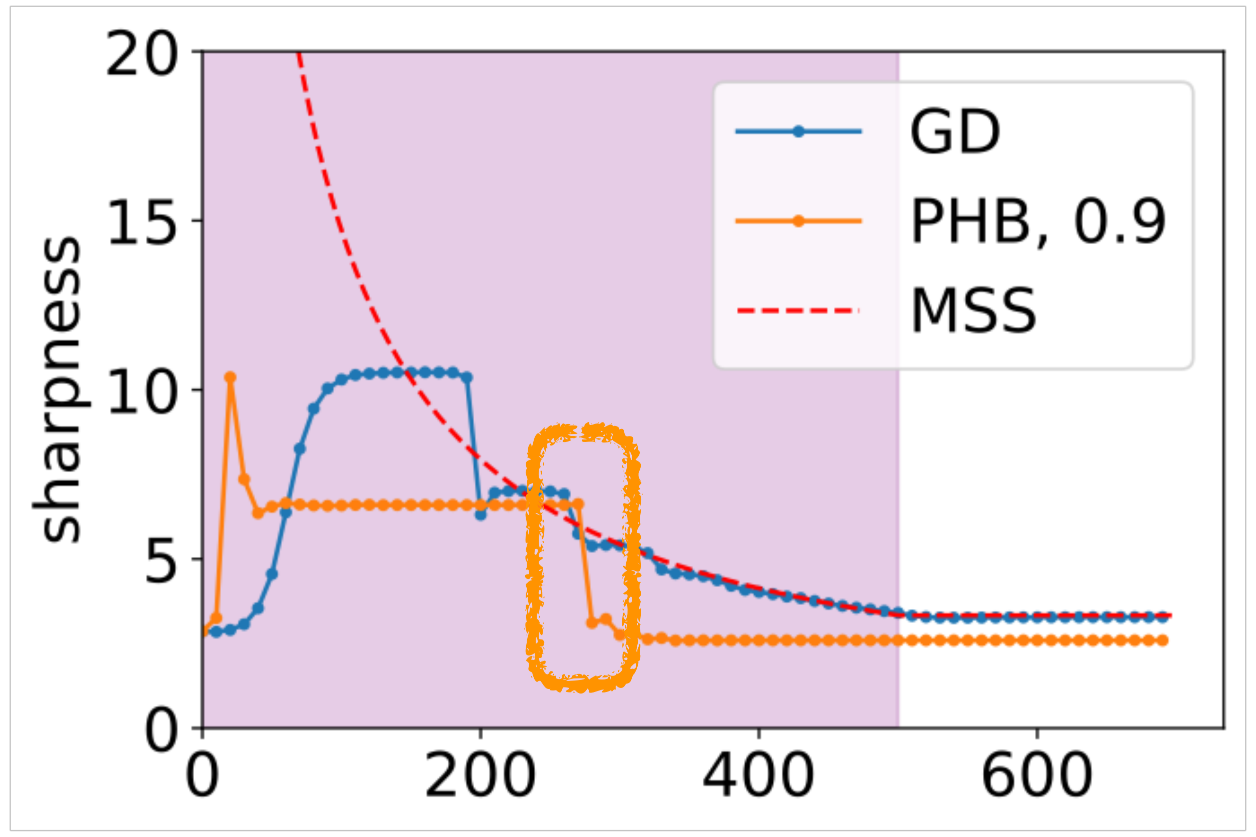


Nonlinear Neural Networks

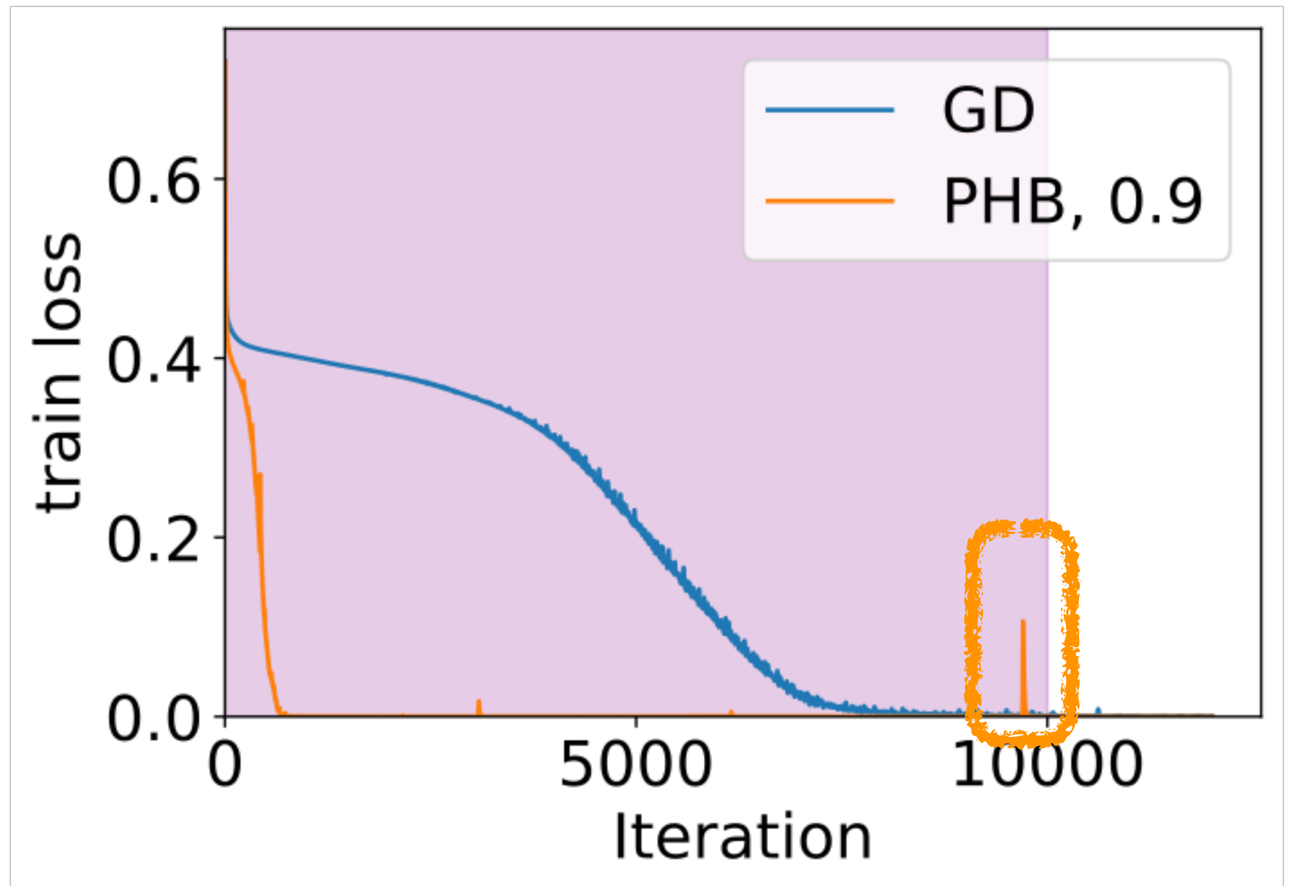
The observations hold for more complex scenarios!

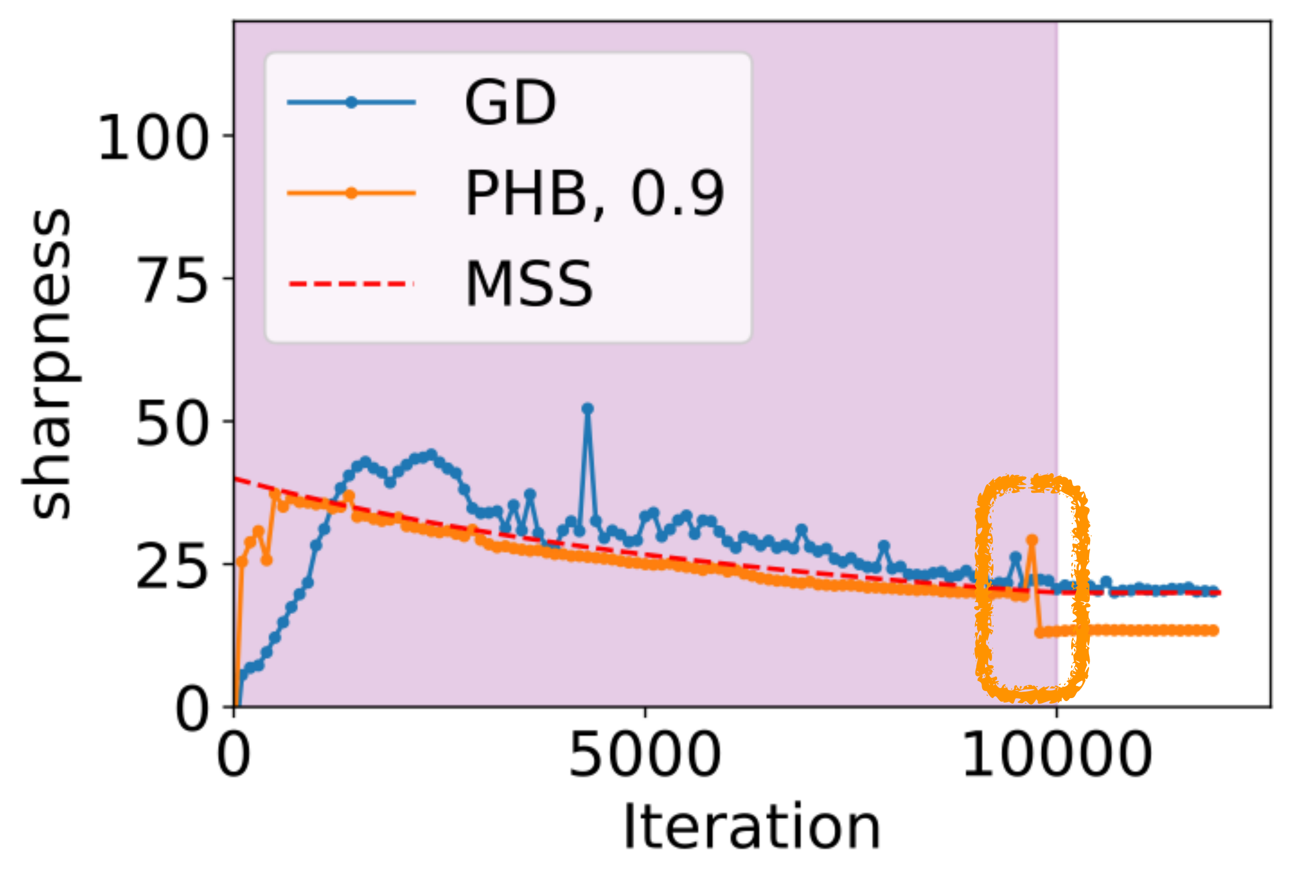
FCN trained on rank-2 (synthetic) dataset:





ResNet20 trained on 1k subset of CIFAR10:





Conclusion & Future Works

Conclusion.

- PHB (with learning rate warmup) induces large catapults, leading to flatter minima
- This is verified over various settings (LDN, toy, nonlinear neural networks)
- The phenomenon is similar to *self-stabilization* effect

Future Works.

- Effect of stochasticity, adaptive momentum? => More extensive <u>experiments</u>
- A complete (or even partial) <u>theoretical characterization</u>

References

A. Lewkowycz, Y. Bahri, E. Dyer, J. Sohl-Dickstein, and G. Gur-Ari. “The large learning rate phase of deep learning: the catapult mechanism.” In *arXiv preprint arXiv:2006.15733,* 2020.

D. Meltzer and J. Liu. “Catapult Dynamics and Phase Transitions in Quadratic Nets.” In *arXiv preprint arXiv:2301.07737* 2023.

L. Zhu, C. Liu, A. Radhakrishnan, and M. Belkin. “Quadratic models for understanding neural network dynamics.” In *arXiv preprint arXiv:2205.11787* 2022.

L. Zhu, C. Liu, A. Radhakrishnan, and M. Belkin. “Catapults in SGD: spikes in the training loss and their impact on generalization through feature learning.” In *arXiv preprint arXiv:2306.04815* 2023.

B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro. “Kernel and Rich Regimes in Overparametrized Models.” In *COLT* 2020.

S. Pesme, L. Pillaud-Vivien, and N. Flammarion. “Implicit Bias of SGD for Diagonal Linear Networks: a Provable Benefit of Stochasticity.” In *NeurIPS* 2021.

M. S. Nacson, K. Ravichandran, N. Srebro, and D. Soudry. “Implicit Bias of the Step Size in Linear Diagonal Neural Networks.” In *ICML* 2022.

S. Pesme and N. Flammarion. “Saddle-to-Saddle Dynamics in Diagonal Linear Networks.” In *NeurIPS* 2023.

M. Even, S. Pesme, S. Gunasekar, and N. Flammarion. “(S)GD over Diagonal Linear Networks: Implicit Regularisation, Large Stepsizes, and Edge of Stability.” In *NeurIPS* 2023.

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 *ICLR* 2021.

A. Damian, E. Nichani, and J. D. Lee. “Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability.” In *ICLR* 2023.

Y. Wang, M. Chen, T. Zhao, and M. Tao. “Large Learning Rate Tames Homogeneity: Convergence and Balancing Effect.” In *ICLR* 2022.

A. Damian, T. Ma, and J. D. Lee. “Label Noise SGD Provably Prefers Flat Global Minimizers.” In *NeurIPS* 2021.

J. Wu, D. Zou, V. Braverman, and Q. Gu. “Direction Matters: On the Implicit Bias of Stochastic Gradient Descent with Moderate Learning Rate.” In *ICLR* 2021.

Y. Li, C. Wei, and T. Ma. “Towards Explaining the Regularization Effect of Initial Large Learning Rate in Training Neural Networks.” In *NeurIPS* 2019.

A. Gotmare, N. S. Keskar, C. Xiong, and R. Socher. “A Closer Look at Deep Learning Heuristics: Learning rate restarts, Warmup and Distillation.” In *ICLR* 2019.

L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. “On the Variance of the Adaptive Learning Rate and Beyond.” In *ICLR* 2020.

Prior Works

Implicit bias of *heavy-ball momentum*

- Small learning rate regime, ODE analysis → stronger (flat) regularizer
  [Ghosh et al., ICLR'23; Wang et al., AAAI'23]
- Binary classification scenarios
  [Jelassi & Li, ICML'22; Wang et al., NeurIPS'22]

We are the *first* to systemically study the dynamics (and implicit bias effect) of momentum in ***large learning rate regime*** for regression setting!