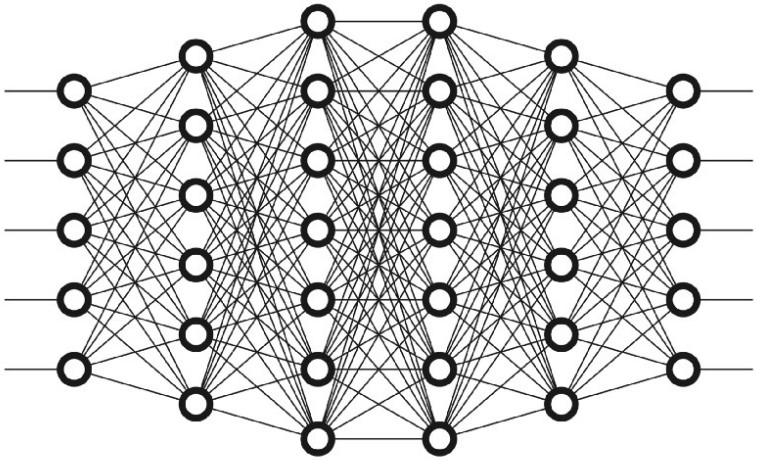
# A Dynamical View on Optimization Algorithms of Overparameterized Neural Networks

Zhiqi Bu  Shiyun Xu  Kan Chen

University of Pennsylvania
{zbu,shiyunxu,kanchen}@upenn.edu

Paper ID: 1332

# Introduction



- Two key factors that lead to the dominating success of neural networks are: network architectures (CNN, ResNet ...) and the optimization (loss functions, optimizers, learning rate scheme ...)
- A convergence theory by many researchers shows that (stochastic) gradient descent provably converges to zero loss (e.g. MSE and cross-entropy) on deep over-parameterized neural networks.

# Introduction

**Previous work on convergence theory:**

- Neural network architecture: FCNN, CNN, ResNet, RNN, GNN, GAN ...
- Loss function: MSE, cross-entropy, non-convex ...
- Optimization algorithm: GD/SGD

**This work:**

- GD with momentums provably converge to zero MSE loss on FCNN;
- Different momentums converge at different rates (exponentially or polynomially);
- Non-convex optimization of neural network can be viewed as some strongly-convex optimization.

*Why does GD converges to global minimum exponentially fast, as if it optimizes a strongly-convex loss, even though the optimization is non-convex?*

# Training dynamics of GD

Now we consider the general training dynamics for an arbitrary neural network $f$, under the MSE loss $L = \frac{1}{2}\|y - f\|^2$.

- Gradient descent (GD): $w(t+1) = w(t) - \eta \frac{\partial L}{\partial w}$
- Gradient flow: $\frac{dw}{dt} = -\frac{\partial L}{\partial w}$
- Dynamics of $f$: $\frac{df}{dt} = \frac{\partial f}{\partial w}\frac{dw}{dt} = -\frac{\partial f}{\partial w}\frac{\partial L}{\partial w} = -\frac{\partial f}{\partial w}(\frac{\partial f}{\partial w})^\top \frac{\partial L}{\partial f}$

We denote the **NTK** matrix $H := \frac{\partial f}{\partial w}(\frac{\partial f}{\partial w})^\top \in \mathbb{R}^{n\times n}$ and rewrite the dynamics:

$$\frac{d(f-y)}{dt} = \frac{df}{dt} = -H\frac{\partial L}{\partial f} = -H(f-y)$$

# Training dynamics of GD

Further denoting the error $\Delta := f - y \in \mathbb{R}^n$, we derive the error dynamics

$$\frac{d(f-y)}{dt} = -H(f-y) \quad \Longrightarrow \quad \dot{\Delta} = -H\Delta$$

Intuitively, treating $H(t)$ as constant and positive definite, we have $\Delta(t) \to 0$ exponentially fast. Hence MSE loss $L(t) = \frac{1}{2}\Delta(t)^\top \Delta(t) \to 0$ exponentially fast.

[Simon Du 2018; Gradient Descent Provably Optimizes Over-parameterized Neural Networks]

# Pseudo-loss and strong convexity

Note that GD only converges exponentially fast on strongly convex loss; it converges $O(1/t)$ on Lipschitz convex loss.

This motivates the definition of a pseudo-loss

$$\hat{L}(t) = \frac{1}{2}\Delta^\top H \Delta.$$

There is a strong resemblance between the non-convex weight dynamics and the strongly-convex error dynamics:

$$\frac{d\mathbf{w}}{dt} = -\frac{\partial L}{\partial \mathbf{w}} \iff \frac{d\Delta}{dt} = -\frac{\partial \hat{L}}{\partial \Delta}$$

$$\frac{d\mathbf{w}}{dt} = -\frac{\partial L}{\partial \mathbf{w}} \iff \frac{d(\boldsymbol{f}-\mathbf{y})}{dt} = -\frac{\partial \hat{L}}{\partial (\boldsymbol{f}-\mathbf{y})}$$

# Optimization with momentum: HBF

We start with the Heavy Ball (HB) method

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \eta \frac{\partial L(\mathbf{w}(k))}{\partial \mathbf{w}(k)} + \beta \left(\mathbf{w}(k) - \mathbf{w}(k-1)\right)$$

Its gradient flow is known as the Heavy Ball with Friction (HBF) system.

$$\ddot{\mathbf{w}}(t) + \sqrt{2\lambda_0}\dot{\mathbf{w}}(t) + \frac{\partial L(\mathbf{w}(t))}{\partial \mathbf{w}(t)} = 0$$

Via the chain rules, we obtain the error dynamics

$$\ddot{\Delta}(t) + \sqrt{2\lambda_0}\dot{\Delta}(t) + \frac{\partial \hat{L}}{\partial \Delta(t)} = 0$$

# Optimization with momentum: HBF

## Theorem (Informal)

*On a two-layer FCNN, suppose we set the width of hidden layer* $m = \Omega\left(\frac{n^6}{\delta^3 \lambda_0^4}\right)$ *and we optimize with Heavy Ball. With high probability at least* $1-\delta$ *over some initialization, we have*

$$L(t) \leq \frac{4}{\lambda_0} \exp\left(-\sqrt{\lambda_0/2} \cdot t\right) \hat{L}(0)$$

[1]

In the proof we use the Lyapunov function by (Siegel, 2019)

$$V(t) := \hat{L} + \tfrac{1}{2}\left\|\sqrt{\tfrac{\lambda_0}{2}}\Delta(t) + \dot{\Delta}(t)\right\|^2 = \tfrac{1}{2}\Delta(t)^\top H(t)\Delta(t) + \tfrac{1}{2}\left\|\sqrt{\tfrac{\lambda_0}{2}}\Delta(t) + \dot{\Delta}(t)\right\|^2.$$

[1] we can show $\sqrt{\lambda_0/2} > \lambda_0$, suggesting HB is faster than GD, at the same order of width.

# Optimization with momentum: NAG

We further study the Nesterov accelerated gradient method (NAG) with time-dependent momentum.

$$\boldsymbol{v}(k+1) = \mathbf{w}(k) - \eta \frac{\partial L(\mathbf{w}(k))}{\partial \mathbf{w}(k)}$$

$$\mathbf{w}(k+1) = \boldsymbol{v}(k+1) + \frac{k-1}{k+\gamma-1}(\boldsymbol{v}(k+1) - \boldsymbol{v}(k))$$

The gradient flow by (Weijie Su, 2014) is

$$\ddot{\mathbf{w}}(t) + \frac{\gamma}{t}\dot{\mathbf{w}}(t) + \frac{\partial L(\mathbf{w}(t))}{\partial \mathbf{w}(t)} = 0.$$

Via the chain rules, we obtain the error dynamics

$$\ddot{\Delta}(t) + \frac{\gamma}{t}\dot{\Delta}(t) + \frac{\partial \hat{L}}{\partial \Delta(t)} = 0$$

For a special NAG (namely `NAG-SC`) with time-independent momentum, the error dynamics is same as HBF and hence enjoys linear convergence.

# On optimization algorithms: NAG

## Theorem (Informal)

*On a two-layer FCNN, suppose we set the width of hidden layer* $m = \Omega\left(\frac{n^{5\alpha/2-4}}{\delta^{3\alpha/2-3}\lambda_0^{3\alpha/2-2}}\right)$ *where* $4 < \alpha \leq \frac{2\gamma}{3}$ *and* $\gamma > 6$*, and we optimize with Nesterov's accelerated gradient. With high probability at least* $1-\delta$ *over some initialization, we have*

$$L(t) \leq A(\alpha, \gamma, \lambda_0) t^{-\alpha} L(0) \tag{1}$$

*where* $A(\alpha, \gamma, \lambda_0)$ *is a constant.*

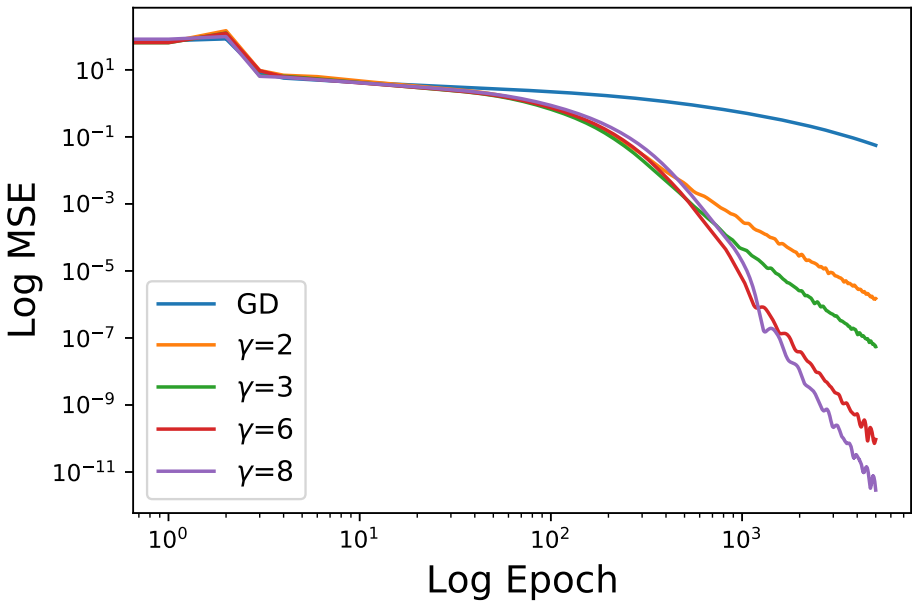
2

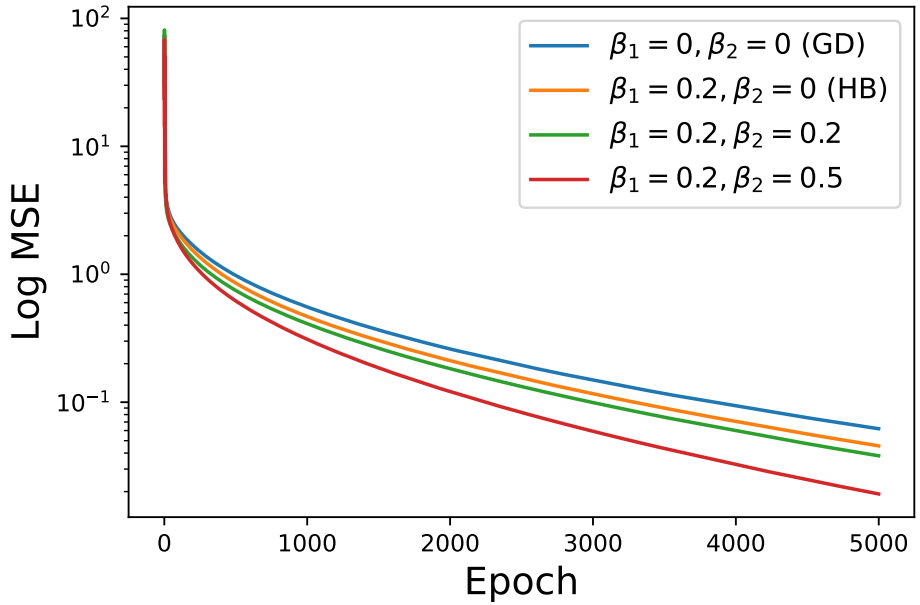
The proof of convergence is much more involved than HB. We use the Lyapunov function by (Weijie Su, 2014)

$$V(t;\alpha,\gamma) := t^{\alpha}\hat{L}(t) + \frac{(2\gamma-\alpha)^2 t^{\alpha-2}}{8}\left\|\Delta(t) + \frac{2t}{2\gamma-\alpha}\dot{\Delta}(t)\right\|^2$$

[2] NAG converges only sublinearly and requires more width in our analysis.

# Illustration of NAG and HBF convergence



# Discussion

## Take home messages

1. Many optimizers besides GD (especially with momentums) provably converge to zero MSE loss.
2. Different momentums converge at different rates with different width requirement.
3. We introduce the pseudo-loss to bridge the classic convex optimization theory to the non-convex neural network optimization, for arbitrary network architecture.

*Extensions:*

1. We briefly discuss extending our work to GD with multiple momentum and the second-order Newton's method.
2. Further extension to optimizers with mini-batch and and adaptive learning rate would be interesting.

# Thank you!