# Provable Accelerated Convergence of Nesterov's Momentum for Deep ReLU Neural Networks

Fangshuo Liao, Anastasios Kyrillidis
*Rice University*

RICE

## Central Question

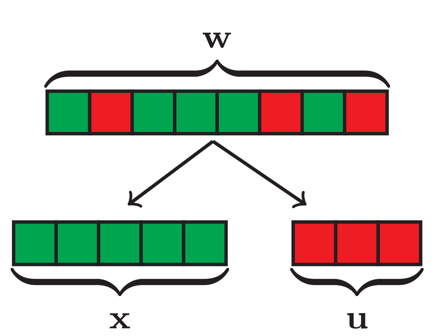
*Does Nesterov's momentum provably converge faster for training neural networks?*

**Our focus**: $\min_{\mathbf{w}} \hat{f}(\mathbf{w})$ where $\hat{f}$ can be non-convex and non-smooth with Nesterov's momentum:

$$\mathbf{w}_{k+1} = \bar{\mathbf{w}}_k - \eta \hat{f}(\bar{\mathbf{w}}_k)$$
$$\bar{\mathbf{w}}_{k+1} = \mathbf{w}_{k+1} + \beta(\mathbf{w}_{k+1} - \mathbf{w}_k) \quad (1)$$

## Parameter Partition

**Definition 1.** *A function $f: \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ is called a partitioned equivalence of $\hat{f}: \mathbb{R}^d \to \mathbb{R}$, if i) $d_1 + d_2 = d$, and ii) there exists a permutation function $\pi: \mathbb{R}^d \to \mathbb{R}^d$ over the parameters of $\hat{f}$, such that $\hat{f}(\mathbf{w}) = f(\mathbf{x}, \mathbf{u})$ if and only if $\pi(\mathbf{w}) = (\mathbf{x}, \mathbf{u})$. We say $(\mathbf{x}, \mathbf{u})$ is a partition of $\mathbf{w}$.*




**Intuition:** Partition the parameters into two sets, with one set having nice properties like strong convexity and smoothness and the other satisfying minimum assumption.

$$\min_{\mathbf{w}\in\mathbb{R}^d} \hat{f}(\mathbf{w}) \equiv \min_{\mathbf{x}\in\mathbb{R}^{d_1}, \mathbf{u}\in\mathbb{R}^{d_2}} f(\mathbf{x}, \mathbf{u})$$

**Nesterov's Momentum on $f(\mathbf{x}, \mathbf{u})$**

$$(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) = (\mathbf{y}_k, \mathbf{v}_k) - \eta \nabla f(\mathbf{y}_k, \mathbf{v}_k)$$
$$(\mathbf{y}_{k+1}, \mathbf{v}_{k+1}) = (\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) + \beta\left((\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) - (\mathbf{x}_k, \mathbf{u}_k)\right) \quad (2)$$

We assume that $f = g \circ h$ with $h: \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}^{\hat{d}}$ (model) and $g: \mathbb{R}^{\hat{d}} \to \mathbb{R}$ ($L_2$-smooth loss).

## Assumptions on u

**Assumption 1.** *$h$ satisfies $G_1$-Lipschitzness with respect to the second part of its parameters:*

$$\|h(\mathbf{x}, \mathbf{u}) - h(\mathbf{x}, \mathbf{v})\|_2 \le G_1 \|\mathbf{u} - \mathbf{v}\|_2,$$
$$\forall \mathbf{x} \in \mathcal{B}^{(1)}_{R_\mathbf{x}};\ \mathbf{u}, \mathbf{v} \in \mathcal{B}^{(2)}_{R_\mathbf{u}}.$$

**Assumption 2.** *The gradient of $f$ with respect to $\mathbf{x}$, namely $\nabla_1 f(\mathbf{x}, \mathbf{u})$, satisfies $G_2$-Lipschitzness with respect to $\mathbf{u}$:*

$$\|\nabla_1 f(\mathbf{x}, \mathbf{u}) - \nabla_1 f(\mathbf{x}, \mathbf{v})\|_2 \le G_2 \|\mathbf{u} - \mathbf{v}\|_2,$$
$$\forall \mathbf{x} \in \mathcal{B}^{(1)}_{R_\mathbf{x}};\ \mathbf{u}, \mathbf{v} \in \mathcal{B}^{(2)}_{R_\mathbf{u}}.$$

## Partial Strong Convexity and Smoothness

**Assumption 3.** *$f$ is $\mu$-strongly convex with $\mu > 0$ with respect to the first part of its parameters:*

$$f(\mathbf{y}, \mathbf{u}) \ge f(\mathbf{x}, \mathbf{u}) + \langle \nabla_1 f(\mathbf{x}, \mathbf{u}), \mathbf{y} - \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|_2^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^{d_1};\ \mathbf{u} \in \mathcal{B}^{(2)}_{R_\mathbf{u}}.$$

**Assumption 4.** *$f$ is $L_1$-smooth with respect to the first part of its parameters:*

$$f(\mathbf{y}, \mathbf{u}) \le f(\mathbf{x}, \mathbf{u}) + \langle \nabla_1 f(\mathbf{x}, \mathbf{u}), \mathbf{y} - \mathbf{x}\rangle + \frac{L_1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^{d_1};\ \mathbf{u} \in \mathcal{B}^{(2)}_{R_\mathbf{u}}.$$

**Assumption 5.** *Minimum values of $f$ restricted to the optimization over $\mathbf{x}$ equal the global minimum value:*

$$\min_{\mathbf{x}\in\mathbb{R}^{d_1}} f(\mathbf{x}, \mathbf{u}) = f^\star := \min_{\mathbf{x}\in\mathbb{R}^{d_1}, \mathbf{u}\in\mathbb{R}^{d_2}} f(\mathbf{x}, \mathbf{u}); \quad \forall \mathbf{u} \in \mathcal{B}^{(2)}_{R_\mathbf{u}}.$$

## How Strong Is Assumption 1-5?

**Theorem 1.** *Let $\tilde{f}$ be $\tilde{\mu}$-strongly convex and $\tilde{L}$-smooth. Then $\tilde{f}$ satisfies Assumptions 1-5 with:*

$$R_\mathbf{x} = R_\mathbf{u} = \infty;\ \mu = \tilde{\mu};\ L_1 = L_2 = \tilde{L};\ G_1 = G_2 = 0.$$

*Also, suppose that Assumption 3, 5 hold. Then, for all $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{u} \in \mathcal{B}^{(2)}_{R_\mathbf{u}}$, we have $\|\nabla f(\mathbf{x}, \mathbf{u})\|_2^2 \ge 2\mu(f(\mathbf{x}, \mathbf{u}) - f^\star)$.*

## Accelerated Convergence under General Assumptions

**Theorem 2** (Gradient Descent). *Suppose that Assumptions 1-5 hold with $G_1^4 \le O\left(\frac{\mu^2}{L_2^2}\right)$ and*

$$R_\mathbf{x} \ge \Omega\left(\eta\kappa\sqrt{L_1}\left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right)^{\frac{1}{2}}\right); \quad R_\mathbf{u} \ge \Omega\left(\eta\kappa G_1\sqrt{L_2}\left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right)^{\frac{1}{2}}\right).$$

*Then there exists constant $c > 0$ such that gradient descent with $\eta = \frac{c}{L_1}$ converges according to:*

$$f(\mathbf{x}_k, \mathbf{u}_k) - f^\star \le \left(1 - \frac{c}{4\kappa}\right)^k \left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right).$$

**Theorem 3** (Nesterov's Momentum). *Let Assumptions 1-5 hold. Consider Nesterov's momentum given by (2) with initialization $\{\mathbf{x}_0, \mathbf{u}_0\} = \{\mathbf{y}_0, \mathbf{v}_0\}$. There exists absolute constants $c, C_1, C_2 > 0$, such that, if $\mu, L_1, L_2, G_1, G_2$ and $R_\mathbf{x}, R_\mathbf{u}$ satisfy:*

$$G_1^4 \le O\left(\frac{\mu^{7/2}}{L_1^{3/2}L_2^3}\right);\ G_1^2 G_2^2 \le O\left(\frac{\mu^{9/2}}{L_1^{\frac{3}{2}}L_2^2}\right)$$
$$R_\mathbf{x} \ge \Omega\left(\frac{L_1^{1/4}L_2^{1/2}}{\mu^{3/4}}\left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right)^{1/2}\right);\ R_\mathbf{u} \ge \Omega\left(\frac{G_1 L_1^{3/4} L_2}{\mu^{7/4}}\left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right)^{1/2}\right), \quad (3)$$

*and, if we choose $\eta = {}^{c}/_{L_1}$, $\beta = {}^{(4\sqrt{\kappa} - \sqrt{c})}/_{(4\sqrt{\kappa} + 7\sqrt{c})}$, then $\mathbf{x}_k, \mathbf{y}_k \in \mathcal{B}^{(1)}_{R_\mathbf{x}}$ and $\mathbf{u}_k, \mathbf{v}_k \in \mathcal{B}^{(2)}_{R_\mathbf{u}}$ for all $k \in \mathbb{N}$, and Nesterov's momentum converges according to:*

$$f(\mathbf{x}_k, \mathbf{u}_k) - f^\star \le 2\left(1 - \frac{c}{4\sqrt{\kappa}}\right)^k \left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right). \quad (4)$$

## Idea of Proof

The core of our proof is to show that $\phi_k$ defined below satisfies $\phi_k \le \left(1 - \frac{c}{4\sqrt{\kappa}}\right)^k \phi_0$:

$$\phi_k = f(\mathbf{x}_k, \mathbf{u}_k) - f^\star + \mathcal{Q}_1\left\|\mathbf{z}_k - \mathbf{x}^\star_{k-1}\right\|_2^2 + \frac{\eta}{8}\left\|\nabla_1 f(\mathbf{y}_{k-1}, \mathbf{v}_{k-1})\right\|_2^2$$

**Difficulty 1: The global minimizer of $f(\mathbf{x}, \mathbf{u})$ may not be unique.** We define a global minimizer $\mathbf{x}^\star(\mathbf{u})$ for each $\mathbf{u}$ and we can show that $\|\mathbf{x}^\star(\mathbf{u}_1) - \mathbf{x}^\star(\mathbf{u}_2)\|_2 \le \frac{G_2}{\mu}\|\mathbf{u}_1 - \mathbf{u}_2\|_2$.

**Difficulty 2: It is not straightforward to control $\nabla_2 f(\mathbf{y}_k, \mathbf{v}_k)$.** We first show that $\|\nabla_2 f(\mathbf{y}_k, \mathbf{v}_k)\|_2^2 \le \frac{G_1^2 L_2}{\mu}\|\nabla_2 f(\mathbf{y}_k, \mathbf{v}_k)\|_2^2$ and bound $\|\nabla_2 f(\mathbf{y}_k, \mathbf{v}_k)\|_2^2$ using a combination of $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2$ and $\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2^2$.

## Neural Networks Satisfy Partial Strong Convexity

$\Lambda$-layer ReLU neural network with layer widths $\{d_\ell\}_{\ell=0}^{\Lambda}$ and activation $\sigma(\mathbf{A})_{ij} = \max\{0, a_{ij}\}$. Let the weight matrix in the $\ell$-th layer be $\mathbf{W}_\ell$. Then, the output of each layer is given by:

$$\mathbf{F}_\ell(\boldsymbol{\theta}) = \begin{cases} \mathbf{X}, & \text{if } \ell = 0; \\ \sigma\left(\mathbf{F}_{\ell-1}(\boldsymbol{\theta})\mathbf{W}_\ell\right), & \text{if } \ell \in [\Lambda - 1]; \\ \mathbf{F}_{\Lambda-1}(\boldsymbol{\theta})\mathbf{W}_\Lambda, & \text{if } \ell = \Lambda. \end{cases} \quad (5)$$

We consider the training of $\mathbf{F}_\Lambda(\boldsymbol{\theta})$ over the MSE loss with data $(\mathbf{X}, \mathbf{Y})$, as in $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{F}_\Lambda(\boldsymbol{\theta}) - \mathbf{Y}\|_F^2$.

**Theorem 4.** *If the width of the network satisfies:*

$$d_\ell = \Theta(m)\ \forall \ell \in [\Lambda - 2];\ d_{\Lambda-1} = \Omega\left(n^{4.5}\max\{n, d^2\}\right),$$

*for some $m \ge \max\{d_0, d_\Lambda\}$, and we initialize the weights according to:*

$$[\mathbf{W}_\ell(0)]_{ij} \sim \mathcal{N}\left(0, d_{\ell-1}^{-1}\right), \quad \forall \ell \in [\Lambda - 1];$$
$$[\mathbf{W}_\Lambda(0)]_{ij} \sim \mathcal{N}\left(0, d_{\Lambda-1}^{-\frac{3}{2}}\right).$$

*Then, with a high probability, there exists a partition $(\mathbf{x}, \mathbf{u})$ of the neural network parameters $\boldsymbol{\theta}$ such that, defining $f = g \circ h$ with $h(\mathbf{x}, \mathbf{u}) = \mathbf{F}_\Lambda(\boldsymbol{\theta})$ and $g(\mathbf{s}) = \frac{1}{2}\|\mathbf{s} - \mathbf{Y}\|_F^2$, we have that $f$ satisfies Assumption 1-5 with $\mu, L_1, L_2, G_1, G_2, R_\mathbf{x}$ and $R_\mathbf{u}$ obeying the condition in (3).*

## Experiments on Additive Models

Let $f = g \circ h$ with $g(\mathbf{s}) = \frac{1}{2}\|\mathbf{s} - \mathbf{y}\|_2^2$ and $h(\mathbf{x}, \mathbf{u}) = \mathbf{A}_1\mathbf{x} + \sigma(\mathbf{A}_2\mathbf{u})$ with $\sigma$ being a $B$-Lipschitz function.

$$f(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\|\mathbf{A}_1\mathbf{x} + \sigma(\mathbf{A}_2\mathbf{u}) - \mathbf{y}\|_2^2$$

We can show that $f$ satisfies Assumptions with small enough $G_1, G_2$ and large enough $R_\mathbf{x}, R_\mathbf{u}$ as long as $\sigma_{\max}(\mathbf{A})$ is small and $\sigma_{\min}(\mathbf{A}_1)$ is large.