On the Width Scaling of Neural Optimizers: Convergence Rates and Intrinsic Width Dependence

Jiajin Li

Sauder School of Business, IAM & CS

University of British Columbia

IOS, March, 2026

Joint work with Yiping Lu (Northwestern).

What this talk is about

From Static to Dynamic:

Analyzing width dependence in convergence rates: stochastic noise induces intrinsic width dependence.

Key Insight from Part I

Definition (L-Smoothness)

Let $f : \Omega \to \mathbb{R}$ be differentiable on a convex set Ω equipped with a norm $\|\cdot\|$. We say that f is *L-smooth* with respect to $\|\cdot\|$ if, for all $\boldsymbol{\Theta}, \boldsymbol{\Theta}' \in \Omega$,

$$\|\nabla f(\boldsymbol{\Theta}) - \nabla f(\boldsymbol{\Theta}')\|_* \leq L\|\boldsymbol{\Theta} - \boldsymbol{\Theta}'\|,$$

where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$.

Width-Independent Smoothness

For feedforward networks with width w, the smoothness constant of the loss function under the mean-normalized geometry,

$$\boxed{(p, \text{mean}) \to (q, \text{mean})}, \quad \text{with } q \geq 2p,$$

is **independent of the width**.

Euclidean Geometry: Why Width Appears

Example (Frobenius smoothness scales with width)

Consider the two-layer network

$$f(\boldsymbol{W}_1, \boldsymbol{W}_2) = \frac{1}{2}(\boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x}))^2,$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{w\times d}$, $\boldsymbol{W}_2 \in \mathbb{R}^{1\times w}$. Assume $\sigma(0) \neq 0$. Then, we have $L_{\mathrm{Fro}} \gtrsim \sigma(0)^2\, w$.

Take $\boldsymbol{W}_1 = 0$, $\boldsymbol{W}_2 = \frac{1}{\sqrt{w}}1_w^\top$, $\boldsymbol{W}_2' = 0$.

Theorem (Informal, (Carmon et al. 2020))

Under the standard L-smoothness assumption, gradient descent finds an ϵ-stationary point with respect to the Frobenius norm in

$$\mathcal{O}\left(\frac{L_{\mathrm{Fro}}\Delta}{\epsilon^2}\right)$$

iterations, where $\Delta := f(\Theta_0) - f^\star$. Moreover, this rate is already optimal for first-order methods.

Euclidean Geometry: Why Width Appears

Example (Frobenius smoothness scales with width)

Consider the two-layer network

$$f(\boldsymbol{W}_1, \boldsymbol{W}_2) = \frac{1}{2}(\boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x}))^2,$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{w\times d}$, $\boldsymbol{W}_2 \in \mathbb{R}^{1\times w}$. Assume $\sigma(0) \neq 0$. Then, we have $L_{\mathrm{Fro}} \gtrsim \sigma(0)^2\, w$.

Take $\boldsymbol{W}_1 = 0$, $\boldsymbol{W}_2 = \frac{1}{\sqrt{w}}1_w^\top$, $\boldsymbol{W}_2' = 0$.

Theorem (Informal, (Carmon et al. 2020))

Under the standard L-smoothness assumption, gradient descent finds an ϵ-stationary point with respect to the Frobenius norm in

$$\mathcal{O}\left(\frac{L_{\mathrm{Fro}}\Delta}{\epsilon^2}\right)$$

iterations, where $\Delta := f(\boldsymbol{\Theta}_0) - f^\star$. Moreover, this rate is already optimal for first-order methods.

Three Computational Geometries

Proposition

Let $\boldsymbol{G} \in \mathbb{R}^{n \times n}$.

- $((2, \mathrm{mean}) \to (2, \mathrm{mean}))$ ***geometry (Muon).***

$$\|\boldsymbol{G}\|_{*,(2,\mathrm{mean})\to(2,\mathrm{mean})} = \|\boldsymbol{G}\|_{\mathrm{nuc}}.$$

- $((p, \mathrm{mean}) \to \infty)$ ***geometry (row normalization).***

$$\|\boldsymbol{G}\|_{*,(p,\mathrm{mean})\to\infty} = n^{-1/p} \sum_{r=1}^{n} \|\boldsymbol{G}_{r,:}\|_p.$$

- $((1, \mathrm{mean}) \to (q, \mathrm{mean}))$ ***geometry (column normalization).***

$$\|\boldsymbol{G}\|_{*,(1,\mathrm{mean})\to(q,\mathrm{mean})} = n^{1/q-1} \sum_{c=1}^{n} \|\boldsymbol{G}_{:,c}\|_{q^*}.$$

Geometry Aware Complexity Bounds

Theorem (Informal, geometry-aware)

Under the L_*-smoothness assumption with respect to a given geometry $\|\cdot\|$, **steepest descent with momentum** *finds an ϵ-stationary point in the dual norm $\|\cdot\|_*$ in*

$$\mathcal{O}\left(\frac{L_*\Delta}{\epsilon^2}\right)$$

iterations, where $\Delta := f(\Theta_0) - f^\star$.

$$\begin{cases} \boldsymbol{D}_k = (1-\alpha)\boldsymbol{D}_{k-1} + \alpha\nabla f(\Theta_k) \\ \boldsymbol{S}_k \in \underset{\|\boldsymbol{S}\|\leq\rho}{\arg\min}\langle \boldsymbol{D}_k, \boldsymbol{S}\rangle \\ \Theta_{k+1} = \Theta_k + \gamma \boldsymbol{S}_k \end{cases}$$

Three Computational Geometries II

Via **norm conversion**, this yields an ϵ-stationary point with respect to the Frobenius norm up to geometry-dependent factors.

Proposition (Relative strength of dual norms)

- ***Muon geometry.***
$$\|\boldsymbol{G}\|_{\mathrm{Fro}} \le \|\boldsymbol{G}\|_{*,(2,\mathrm{mean})\to(2,\mathrm{mean})} = \|\boldsymbol{G}\|_{\mathrm{nuc}}.$$
- ***Row-normalization geometry.*** *For any* $1 \le p \le \infty$,
$$\|\boldsymbol{G}\|_{\mathrm{Fro}} \le n^{\max\{\frac{1}{p},\frac{1}{2}\}} \|\boldsymbol{G}\|_{*,(p,\mathrm{mean})\to\infty}.$$
- ***Column-normalization geometry.*** *For any* $1 \le q \le \infty$,
$$\|\boldsymbol{G}\|_{\mathrm{Fro}} \le n^{\max\{1-\frac{1}{q},\frac{1}{2}\}} \|\boldsymbol{G}\|_{*,(1,\mathrm{mean})\to(q,\mathrm{mean})}.$$

Chosen Geometry

Theorem (Informal, geometry-aware)

Under the L_*-smoothness assumption with respect to a geometry $\|\cdot\|$, **steepest descent with momentum** *finds an ϵ-stationary point in the dual norm $\|\cdot\|_*$ in*

$$\mathcal{O}\left(\frac{L_*\Delta}{\epsilon^2}\right)$$

iterations, where $\Delta := f(\mathbf{\Theta}_0) - f^\star$.

For **Muon**, width dependence arises from L_* (intrinsic), whereas for **row/column normalization**, it arises from norm conversion (extrinsic).

Choosing geometry $\{2 \to 2,\ p \geq 2,\ q = 2\}$ yields at least a $\sqrt{w}$ improvement.

Chosen Geometry

Theorem (Informal, geometry-aware)

Under the L_*-smoothness assumption with respect to a geometry $\|\cdot\|$, **steepest descent with momentum** *finds an ϵ-stationary point in the dual norm $\|\cdot\|_*$ in*

$$\mathcal{O}\left(\frac{L_*\Delta}{\epsilon^2}\right)$$

iterations, where $\Delta := f(\mathbf{\Theta}_0) - f^\star$.

For **Muon**, width dependence arises from L_* (intrinsic), whereas for **row/column normalization**, it arises from norm conversion (extrinsic).

Choosing geometry $\{2\to 2,\ p \geq 2,\ q = 2\}$ yields at least a $\sqrt{w}$ improvement.

Finite Sum Optimization Problem

Consider the finite-sum optimization problem

$$\min_{\boldsymbol{\Theta}\in\Omega_C} F(\boldsymbol{\Theta}) \;\coloneqq\; \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{\Theta}, \boldsymbol{x}_n), \tag{P}$$

where $\boldsymbol{\Theta}$ denotes the model parameter, $\{\boldsymbol{x}_n\}_{n=1}^N$ is the given dataset, and $f(\boldsymbol{\Theta}, \boldsymbol{x}_n)$ is the sample-wise loss associated with the data point $\boldsymbol{x}_n$.

Assumption (Unbiasedness and bounded variance)

For any Θ, let $g(\Theta, \xi)$ be a stochastic gradient estimator. We assume:

- ***Unbiasedness:***

$$\mathbb{E}_\xi\big[g(\Theta,\xi)\big] = \nabla F(\Theta).$$

- ***Bounded variance:***

$$\mathbb{E}_\xi\big[\|g(\Theta,\xi) - \nabla F(\Theta)\|_{\mathrm{Fro}}{}^2\big] \le \sigma_{\mathrm{Fro}}^2.$$

Finite Sum Optimization Problem

Consider the finite-sum optimization problem

$$\min_{\boldsymbol{\Theta}\in\Omega_C} F(\boldsymbol{\Theta}) \;:=\; \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{\Theta}, \boldsymbol{x}_n), \tag{P}$$

where $\boldsymbol{\Theta}$ denotes the model parameter, $\{\boldsymbol{x}_n\}_{n=1}^N$ is the given dataset, and $f(\boldsymbol{\Theta}, \boldsymbol{x}_n)$ is the sample-wise loss associated with the data point $\boldsymbol{x}_n$.

Assumption (Unbiasedness and bounded variance)

For any $\boldsymbol{\Theta}$, let $g(\boldsymbol{\Theta}, \xi)$ be a stochastic gradient estimator. We assume:

- ***Unbiasedness:***

$$\mathbb{E}_\xi\big[g(\boldsymbol{\Theta}, \xi)\big] = \nabla F(\boldsymbol{\Theta}).$$

- ***Bounded variance:***

$$\mathbb{E}_\xi\big[\|g(\boldsymbol{\Theta}, \xi) - \nabla F(\boldsymbol{\Theta})\|_{\mathrm{Fro}}{}^2\big] \le \sigma_{\mathrm{Fro}}^2.$$

Finite Sum Optimization Problem

Consider the finite-sum optimization problem

$$\min_{\Theta \in \Omega_C} F(\Theta) \;:=\; \frac{1}{N} \sum_{n=1}^{N} f(\Theta, \boldsymbol{x}_n), \tag{P}$$

where Θ denotes the model parameter, $\{\boldsymbol{x}_n\}_{n=1}^N$ is the given dataset, and $f(\Theta, \boldsymbol{x}_n)$ is the sample-wise loss associated with the data point $\boldsymbol{x}_n$.

Theorem (Informal, (Arjevani et al. 2023))

Under the standard L-smoothness and noise assumptions, stochastic gradient descent finds an ϵ-stationary point with respect to the Frobenius norm in

$$\mathcal{O}\Big(\frac{L_{\mathrm{Fro}}\sigma_{\mathrm{Fro}}^2\Delta}{\epsilon^4}\Big)$$

iterations, where $\Delta := f(\Theta_0) - f^\star$. Moreover, this rate is already optimal for stochastic first-order methods.

Finite Sum Optimization Problem

Consider the finite-sum optimization problem

$$\min_{\boldsymbol{\Theta}\in\Omega_C} F(\boldsymbol{\Theta}) \;:=\; \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{\Theta}, \boldsymbol{x}_n), \tag{P}$$

where $\boldsymbol{\Theta}$ denotes the model parameter, $\{\boldsymbol{x}_n\}_{n=1}^N$ is the given dataset, and $f(\boldsymbol{\Theta}, \boldsymbol{x}_n)$ is the sample-wise loss associated with the data point $\boldsymbol{x}_n$.

Theorem (Informal, (Arjevani et al. 2023))

Under the standard L-smoothness and noise assumptions, stochastic gradient descent finds an ϵ-stationary point with respect to the Frobenius norm in

$$\mathcal{O}\left(\frac{L_{\mathrm{Fro}}\sigma_{\mathrm{Fro}}^2\Delta}{\epsilon^4}\right)$$

iterations, where $\Delta := f(\boldsymbol{\Theta}_0) - f^\star$. Moreover, this rate is already optimal for stochastic first-order methods.

Euclidean Geometry: Why Width Appears Again

Example (A two-layer network whose stochastic gradient variance scales with width)

Consider the two-layer scalar-output network

$$f_i(\boldsymbol{W}_1, \boldsymbol{W}_2) := \frac{1}{2}(\boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x}) - a_i)^2,$$

where $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{W}_1 \in \mathbb{R}^{w\times d}$, $\boldsymbol{W}_2 \in \mathbb{R}^{1\times w}$, and σ acts entrywise. Let $a_1 = 1$ and $a_2 = -1$, and define the finite-sum objective

$$F(\boldsymbol{W}_1, \boldsymbol{W}_2) = \frac{1}{2}\big(f_1(\boldsymbol{W}_1, \boldsymbol{W}_2) + f_2(\boldsymbol{W}_1, \boldsymbol{W}_2)\big).$$

Assume that $\sigma(0) \neq 0$. Then, under uniform sampling $i \sim \mathrm{Unif}\{1, 2\}$ with batch size $B = 1$, the stochastic gradient satisfies

$$\mathbb{E}\big[\|\nabla f_i(\boldsymbol{W}_1, \boldsymbol{W}_2) - \nabla F(\boldsymbol{W}_1, \boldsymbol{W}_2)\|_{\mathrm{Fro}}^2\big] \geq \sigma(0)^2\, w.$$

Width-Independent Variance for Chosen Geometries

Theorem (Ours)

Suppose that (1) $1 \le p \le q < \infty$*; (2) the input is bounded; (3) both* σ *and* $\mathcal{L}$ *have bounded second derivatives. Then, for any finite dataset* $\{\boldsymbol{x}_i\}_{i=1}^N$*,*

$$\|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}; \boldsymbol{x}_i) - \nabla_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}; \boldsymbol{x}_j)\|_*^2 \le \sigma_*^2, \qquad \forall i, j \in [N],$$

where σ_* **is independent of the width** w*.*

Implication (stochastic oracle):

Let $\xi \sim \mathrm{Unif}([N])$ and $g(\Theta, \xi) = \nabla f(\Theta; x_\xi)$. Then

$$\mathbb{E}_\xi \left[\|g(\Theta, \xi) - \nabla F(\Theta)\|_*^2 \right] \le \sigma_*^2.$$

Is dimension free complexity bound possible?

Width-Independent Variance for Chosen Geometries

Theorem (Ours)

Suppose that (1) $1 \le p \le q < \infty$*; (2) the input is bounded; (3) both* σ *and* $\mathcal{L}$ *have bounded second derivatives. Then, for any finite dataset* $\{\boldsymbol{x}_i\}_{i=1}^N$*,*

$$\|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}; \boldsymbol{x}_i) - \nabla_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}; \boldsymbol{x}_j)\|_*^2 \le \sigma_*^2, \qquad \forall i, j \in [N],$$

where σ_* **is independent of the width** w*.*

Implication (stochastic oracle):

Let $\xi \sim \mathrm{Unif}([N])$ and $g(\boldsymbol{\Theta}, \xi) = \nabla f(\boldsymbol{\Theta}; \boldsymbol{x}_\xi)$. Then

$$\mathbb{E}_\xi \left[\|g(\boldsymbol{\Theta}, \xi) - \nabla F(\boldsymbol{\Theta})\|_*{}^2 \right] \le \sigma_*^2.$$

Is dimension free complexity bound possible?

Width-Independent Variance for Chosen Geometries

Theorem (Ours)

Suppose that (1) $1 \le p \le q < \infty$*; (2) the input is bounded; (3) both* σ *and* $\mathcal{L}$ *have bounded second derivatives. Then, for any finite dataset* $\{\boldsymbol{x}_i\}_{i=1}^N$*,*

$$\|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}; \boldsymbol{x}_i) - \nabla_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}; \boldsymbol{x}_j)\|_*^2 \le \sigma_*^2, \qquad \forall i, j \in [N],$$

where σ_* **is independent of the width** w*.*

Implication (stochastic oracle):

Let $\xi \sim \mathrm{Unif}([N])$ and $g(\boldsymbol{\Theta}, \xi) = \nabla f(\boldsymbol{\Theta}; \boldsymbol{x}_\xi)$. Then

$$\mathbb{E}_\xi \left[\|g(\boldsymbol{\Theta}, \xi) - \nabla F(\boldsymbol{\Theta})\|_*{}^2 \right] \le \sigma_*^2.$$

Is dimension free complexity bound possible?

Lower Bound: No Dimension-Free Complexity

All chosen geometries can be reduced to a scaled (ℓ_∞, ℓ_1) vector geometry.

Theorem (Ours)

If $d > 2T$, then for any randomized T-query algorithm A,

$$\sup_{(f,\mathcal{O})\in\mathfrak{F}_d} \mathbb{E}\left[\|\nabla f(\widehat{x})\|_1\right] > \epsilon.$$

$$d > 2T \quad \Longrightarrow \quad T = \Omega(d).$$

Linear dependence on dimension is unavoidable

Lower Bound: No Dimension-Free Complexity

All chosen geometries can be reduced to a scaled (ℓ_∞, ℓ_1) vector geometry.

Theorem (Ours)

If $d > 2T$, then for any randomized T-query algorithm A,

$$\sup_{(f,\mathcal{O})\in\mathfrak{F}_d} \mathbb{E}\left[\|\nabla f(\widehat{x})\|_1\right] > \epsilon.$$

$$d > 2T \qquad \Longrightarrow \qquad T = \Omega(d).$$

Linear dependence on dimension is unavoidable

Upper Bound: Stochastic Steepest Descent with Momentum

If we choose geometry $\{2 \to 2,\ p \geq 2,\ q = 2\}$, we will have

Theorem ((Shen et al. 2025), (Pethick et al. 2025))

Under the L_*-smoothness and noise assumptions with respect to a given geometry $\|\cdot\|$, **stochastic steepest descent with momentum** *finds an ϵ-stationary point in the dual norm $\|\cdot\|_*$ in*

$$\mathcal{O}\left(\frac{w L_* \sigma_*^2 \Delta}{\epsilon^4}\right)$$

iterations, where $\Delta := f(\boldsymbol{\Theta}_0) - f^\star$.

Width dependence is not intrinsic to the function class, but arises from stochastic noise.

Thank you for your listening!

Any questions?