Beyond Uniform Smoothness: A Stopped Analysis of Adaptive SGD

Speaker: Matthew Faw

Authors: F*, Litu Rout*, Constantine Caramanis, and Sanjay Shakkottai

The University of Texas at Austin

*=Equal contribution

Problem Setup

Find a first-order stationary point of a non-convex, L_0-smooth function F:

$$||\nabla F(x) - \nabla F(y)|| \leq L_0 ||x - y|| \quad \forall\, x, y$$

When F is twice-differentiable, equivalent to:

$$||\nabla^2 F(x)|| \leq L_0 \ \forall x$$

Problem Setup

Find a first-order stationary point of a non-convex, (L_0, L_1)-smooth[1] function F:

$$||\nabla F(x) - \nabla F(y)|| \leq (L_0 + L_1||\nabla F(y)||)||x - y|| \quad \forall\ ||x - y|| \leq 1/L_1$$

When F is twice-differentiable, ≈ equivalent to:

$$||\nabla^2 F(x)|| \leq L_0 + L_1||\nabla F(x)|| \ \ \forall x$$

[1]Defined by [Zhang-He-Sra-Jadbabaie’20, Zhang-Jin-Fang-Wang’20]

Problem Setup

Find a first-order stationary point of a non-convex, (L_0, L_1)-smooth[1] function F:

$$||\nabla F(x) - \nabla F(y)|| \leq (L_0 + L_1||\nabla F(y)||)||x - y|| \quad \forall\, ||x - y|| \leq 1/L_1$$

When F is twice-differentiable, ≈ equivalent to:

$$||\nabla^2 F(x)|| \leq L_0 + L_1||\nabla F(x)|| \ \ \forall x$$

- Standard L-smoothness is equivalent to $(L, 0)$-smoothness

[1]Defined by [Zhang-He-Sra-Jadbabaie'20, Zhang-Jin-Fang-Wang'20]

Problem Setup

Find a first-order stationary point of a non-convex, (L_0, L_1)-smooth[1] function F:

$$||\nabla F(x) - \nabla F(y)|| \leq (L_0 + L_1||\nabla F(y)||)||x - y|| \quad \forall\, ||x - y|| \leq 1/L_1$$

When F is twice-differentiable, ≈ equivalent to:

$$||\nabla^2 F(x)|| \leq L_0 + L_1||\nabla F(x)|| \;\; \forall x$$

- Standard L-smoothness is equivalent to $(L, 0)$-smoothness
- Also captures a wide class of functions which are not uniformly smooth, e.g.:
  - $F(x) = x^c$ for $c > 2$ – $(c(c-1), c-1)$-smooth
  - $F(x) = e^{c'x}$ for $c > 0$ – $(0, c')$-smooth

[1]Defined by [Zhang-He-Sra-Jadbabaie’20, Zhang-Jin-Fang-Wang’20]

Prior work on non-uniform smoothness

Problem Setup

Find a first-order stationary point of a non-convex, (L_0, L_1)-smooth function F:

$$||\nabla F(x) - \nabla F(y)|| \leq (L_0 + L_1||\nabla F(y)||)||x - y|| \quad \forall\ ||x - y|| \leq 1/L_1$$

Normalized/Clipped SGD

$$w_{t+1} = w_t - \eta_t g_t \qquad \eta_t = \begin{cases} \frac{\eta}{\gamma + ||g_t||} & \text{Normalized SGD} \\ \frac{\eta}{\max\{\gamma, ||g_t||\}} & \text{Clipped SGD} \end{cases}$$

Problem Setup

Find a first-order stationary point of a non-convex, (L_0, L_1)-smooth function F:

$$||\nabla F(w_{t+1}) - \nabla F(w_t)|| \leq (L_0 + L_1||\nabla F(w_t)||)||w_{t+1} - w_t|| \quad \forall\ ||w_{t+1} - w_t|| \leq 1/L_1$$

To *deterministically* satisfy

Normalized/Clipped SGD

$$w_{t+1} = w_t - \eta_t g_t \qquad \eta_t = \begin{cases} \frac{\eta}{\gamma + ||g_t||} & \text{Normalized SGD} \\ \frac{\eta}{\max\{\gamma, ||g_t||\}} & \text{Clipped SGD} \end{cases}$$

Problem Setup

Find a first-order stationary point of a non-convex, (L_0, L_1)-smooth function F:

$$||\nabla F(w_{t+1}) - \nabla F(w_t)|| \leq (L_0 + L_1||\nabla F(w_t)||)||w_{t+1} - w_t|| \quad \forall\, ||w_{t+1} - w_t|| \leq 1/L_1$$

To *deterministically* satisfy

Normalized/Clipped SGD

$$w_{t+1} = w_t - \eta_t g_t \qquad \eta_t = \begin{cases} \frac{\eta}{\gamma + ||g_t||} & \text{Normalized SGD} \\ \frac{\eta}{\max\{\gamma, ||g_t||\}} & \text{Clipped SGD} \end{cases}$$

Prior work[1] established $\min_t ||\nabla F(w_t)||^2 = O(1/\sqrt{T})$ convergence rate assuming:

- $\mathbb{E}[g] = \nabla F(w)$ (unbiased stochastic gradient)
- $\sup_w ||g - \nabla F(w)||^2 \leq \sigma_0^2$ (bounded *noise support*)

[1] [Zhang-He-Sra-Jadbabaie'20, Zhang-Jin-Fang-Wang'20, Crawshaw-Liu-Orabona-Zhang-Zhuang'22,…]

Problem Setup

Find a first-order stationary point of a non-convex, (L_0, L_1)-smooth function F:

$$||\nabla F(w_{t+1}) - \nabla F(w_t)|| \leq (L_0 + L_1||\nabla F(w_t)||)||w_{t+1} - w_t|| \quad \forall\ ||w_{t+1} - w_t|| \leq 1/L_1$$

To *deterministically* satisfy

Normalized/Clipped SGD

$$w_{t+1} = w_t - \eta_t g_t \qquad \eta_t = \begin{cases} \frac{\eta}{\gamma + ||g_t||} & \text{Normalized SGD} \\ \frac{\eta}{\max\{\gamma, ||g_t||\}} & \text{Clipped SGD} \end{cases}$$

Prior work[2] established $\min_t ||\nabla F(w_t)||^2 = O(1/\sqrt{T})$ convergence rate assuming:

- $\mathbb{E}[g] = \nabla F(w)$ (unbiased stochastic gradient)
- $\sup_w ||g - \nabla F(w)||^2 \leq \sigma_0^2$ (bounded *noise support*)

Significantly stronger assumption than is needed in L-smooth setting

[2] [Zhang-He-Sra-Jadbabaie'20, Zhang-Jin-Fang-Wang'20, Crawshaw-Liu-Orabona-Zhang-Zhuang'22,…]

Problem Setup

Find a first-order stationary point of a non-convex, $(L_0, 0)$-smooth function F:

$$||\nabla F(x) - \nabla F(y)|| \leq L_0||x - y|| \quad \forall\, x, y$$

AdaGrad-Norm

$$w_{t+1} = w_t - \eta_t g_t$$

$$\eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t} ||g_s||^2}}$$

Prior work[3] established $\min_t ||\nabla F(w_t)||^2 = O(1/\sqrt{T})$ convergence rate assuming:

- $\mathbb{E}[g] = \nabla F(w)$ (unbiased stochastic gradient)
- $\mathbb{E}[||g - \nabla F(w)||^2] \leq \sigma_0^2 + \sigma_1^2||\nabla F(w)||^2$ (*affine* variance)

[3] [Li-Orabona’19,’20; Ward-Wu-Bottou’19; Kavis-Levy-Cevher’22; **F**-Tziotis-Caramanis-Mokhtari-Shakkottai-Ward’22]

Problem Setup

Find a first-order stationary point of a non-convex, $(L_0, 0)$-smooth function F:

$$||\nabla F(x) - \nabla F(y)|| \leq L_0 ||x - y|| \quad \forall\, x, y$$

AdaGrad-Norm

$$w_{t+1} = w_t - \eta_t g_t$$

No tuning w.r.t. L_0!

$$\eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t} ||g_s||^2}}$$

Prior work[3] established $\min_t ||\nabla F(w_t)||^2 = O(1/\sqrt{T})$ convergence rate assuming:

- $\mathbb{E}[g] = \nabla F(w)$ (unbiased stochastic gradient)
- $\mathbb{E}[||g - \nabla F(w)||^2] \leq \sigma_0^2 + \sigma_1^2 ||\nabla F(w)||^2$ (*affine* variance)

[3] [Li-Orabona'19,'20; Ward-Wu-Bottou'19; Kavis-Levy-Cevher'22; **F**-Tziotis-Caramanis-Mokhtari-Shakkottai-Ward'22]

Problem Setup

Find a first-order stationary point of a non-convex, $(L_0, 0)$-smooth function F:

$$||\nabla F(x) - \nabla F(y)|| \leq L_0||x - y|| \quad \forall\, x, y$$

AdaGrad-Norm

$$w_{t+1} = w_t - \eta_t g_t$$

$$\eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t} ||g_s||^2}}$$

No tuning w.r.t. L_0!

Prior work[3] established $\min_t ||\nabla F(w_t)||^2 = O(1/\sqrt{T})$ convergence rate assuming:

- $\mathbb{E}[g] = \nabla F(w)$ (unbiased stochastic gradient)
- $\mathbb{E}[||g - \nabla F(w)||^2] \leq \sigma_0^2 + \sigma_1^2||\nabla F(w)||^2$ (*affine* variance)

Analysis *heavily* relies on the L_0-smoothness assumption

[3] [Li-Orabona'19,'20; Ward-Wu-Bottou'19; Kavis-Levy-Cevher'22; **F**-Tziotis-Caramanis-Mokhtari-Shakkottai-Ward'22]

Problem Setup

Find a first-order stationary point of a non-convex, (L_0, L_1)-smooth function F:

$$||\nabla F(x) - \nabla F(y)|| \leq (L_0 + L_1||\nabla F(y)||)||x - y|| \quad \forall\ ||x - y|| \leq 1/L_1$$

AdaGrad-Norm

$$w_{t+1} = w_t - \eta_t g_t$$

$$\eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t} ||g_s||^2}}$$

No tuning w.r.t. L_0!

Question

Given that AdaGrad-Norm adapts to the smoothness parameter L_0 automatically…

Is it possible to prove that AdaGrad-Norm converges at rate $\tilde{O}(1/\sqrt{T})$ under:

- (L_0, L_1)-smoothness
- *Affine* variance

Problem Setup

Find a first-order stationary point of a non-convex, (L_0, L_1)-smooth function F:

$$||\nabla F(x) - \nabla F(y)|| \leq (L_0 + L_1||\nabla F(y)||)||x - y|| \quad \forall\ ||x - y|| \leq 1/L_1$$

AdaGrad-Norm

$$w_{t+1} = w_t - \eta_t g_t$$

$$\eta_t = \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^t ||g_s||^2}}$$

No tuning w.r.t. L_0!

Question

Given that AdaGrad-Norm adapts to the smoothness parameter L_0 automatically…

Is it possible to prove that AdaGrad-Norm converges at rate $\tilde{O}(1/\sqrt{T})$ under:

- (L_0, L_1)-smoothness
- *Affine* variance

Yes!

Challenges for Adaptive Methods

Overcoming the challenges of **adaptive** step sizes

- **Challenge 1**: Bias + affine variance
  - Step size η_t depends on past and **current stochastic gradients**.

Descent direction $-\eta_t g_t$ is **biased!**

Descent direction $-\eta_t g_t$ is **biased!**

Overcoming the challenges of **adaptive** step sizes

- **Challenge 1**: Bias + affine variance
  - Step size η_t depends on past and **current stochastic gradients**.
  - $\Longrightarrow$ Obtaining a useful descent lemma from smoothness becomes challenging

$$\eta_t \|\nabla F(w_t)\|^2 \leq F(w_t) - F(w_{t+1}) + \underbrace{\eta_t \langle \nabla F(w_t), \nabla F(w_t) - g_t \rangle}_{\text{Not mean-zero!}} + \frac{(L_0 + L_1 \|\nabla F(w_t)\|)\eta_t^2}{2} \|g_t\|^2$$

Descent direction $-\eta_t g_t$ is **biased!**

Overcoming the challenges of **adaptive** step sizes

- **Challenge 1**: Bias + affine variance
  - Step size η_t depends on past and **current stochastic gradients**.
  - $\Longrightarrow$ Obtaining a useful descent lemma from smoothness becomes challenging

$$\eta_t\|\nabla F(w_t)\|^2 \leq F(w_t) - F(w_{t+1}) \underbrace{+\, \eta_t\langle \nabla F(w_t), \nabla F(w_t) - g_t\rangle}_{\text{Not mean-zero!}} + \frac{(L_0 + L_1\|\nabla F(w_t)\|)\eta_t^2}{2}\|g_t\|^2$$

  - Especially challenging under **affine variance**

$$\tilde{\eta}_t(1 - \sigma_1 \cdot bias_t)\|\nabla F(w_t)\|^2 \leq \mathbb{E}_t[F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t[\eta_t^2\|g_t\|^2]$$

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c + \sum_{s<t}\|g_s\|^2 + c'\|\nabla F(w_t)\|^2}} \quad \text{and} \quad bias_t = \sqrt{E_t\left[\frac{\|g_t\|^2}{b_0^2 + \sum_{s=1}^{t}\|g_s\|^2}\right]}$$

Descent direction $-\eta_t g_t$ is **biased!**

Overcoming the challenges of **adaptive** step sizes

- **Challenge 1**: Bias + affine variance
  - Step size η_t depends on past and **current stochastic gradients**.
  - Especially challenging under **affine variance**

$$\tilde{\eta}_t(1 - \sigma_1 \cdot bias_t)\|\nabla F(w_t)\|^2 \leq \mathbb{E}_t[F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t[\eta_t^2 \|g_t\|^2]$$

Step-size "proxy"

Lower bound for $\mathbb{E}_t[\eta_t]$

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c + \sum_{s<t} ||g_s||^2 + c' ||\nabla F(w_t)||^2}} \quad \text{and} \quad bias_t = \sqrt{E_t\left[\frac{||g_t||^2}{b_0^2 + \sum_{s=1}^{t} ||g_s||^2}\right]}$$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 1**: Bias + affine variance
  - Step size η_t depends on past and **current stochastic gradients**.
  - Especially challenging under **affine variance**

Descent direction $-\eta_t g_t$ is **biased!**

$$\tilde{\eta}_t \underbrace{(1 - \sigma_1 \cdot bias_t)}_{\text{Possibly } negative} \|\nabla F(w_t)\|^2 \leq \mathbb{E}_t[F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t[\eta_t^2 \|g_t\|^2]$$

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c + \sum_{s<t} ||g_s||^2 + c' ||\nabla F(w_t)||^2}} \quad \text{and} \quad bias_t = \sqrt{E_t\left[\frac{||g_t||^2}{b_0^2 + \sum_{s=1}^{t} ||g_s||^2}\right]}$$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 1**: Bias + affine variance
  - Step size η_t depends on past and **current stochastic gradients**.
  - Especially challenging under **affine variance**

Descent direction $-\eta_t g_t$ is **biased!**

$$\underbrace{\tilde{\eta}_t (1 - \sigma_1 \cdot bias_t)}_{\geq \frac{1}{2}} \|\nabla F(w_t)\|^2 \leq \mathbb{E}_t[F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t[\eta_t^2 \|g_t\|^2]$$

Key Idea 1:
Focus on the ***"good" times***
when bound is non-vacuous

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c + \sum_{s<t} ||g_s||^2 + c' ||\nabla F(w_t)||^2}} \quad \text{and} \quad bias_t = \sqrt{E_t \left[\frac{||g_t||^2}{b_0^2 + \sum_{s=1}^{t} ||g_s||^2} \right]}$$

Challenges for Adaptive Methods

Overcoming the challenges of **adaptive** step sizes

Descent direction $-\eta_t g_t$ is **biased!**

- **Challenge 1**: Bias + affine variance
  - Step size η_t depends on past and **current stochastic gradients**.
  - Especially challenging under **affine variance**

    $$\tilde{\eta}_t(1 - \sigma_1 \cdot bias_t)\|\nabla F(w_t)\|^2 \leq \mathbb{E}_t[F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t[\eta_t^2\|g_t\|^2]$$

  - Most times are (typically) "good" $\Longrightarrow$ descent inequality (roughly) of the form:

    $$\mathbb{E}\left[\sum_{t\leq T} \tilde{\eta}_t||\nabla F(w_t)||^2\right] \leq F(w_0) - F^* + c \text{ poly} \log(T)$$

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c+\sum_{s<t}||g_s||^2+c'||\nabla F(w_t)||^2}} \quad \text{and} \quad bias_t = \sqrt{E_t\left[\frac{||g_t||^2}{b_0^2+\sum_{s=1}^{t}||g_s||^2}\right]}$$

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c + \sum_{s<t} ||g_s||^2 + c'||\nabla F(w_t)||^2}}$$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - How to obtain a convergence rate from the following descent inequality?

$$\mathbb{E}[\sum_{t \leq T} \tilde{\eta}_t ||\nabla F(w_t)||^2] \leq F(w_0) - F^* + c \text{ poly} \log(T)$$

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c + \sum_{s<t} ||g_s||^2 + c'||\nabla F(w_t)||^2}}$$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - How to obtain a convergence rate from the following descent inequality?
    $$\mathbb{E}[\textstyle\sum_{t\le T} \tilde{\eta}_t ||\nabla F(w_t)||^2] \le F(w_0) - F^* + c \text{ poly} \log(T)$$
  - Would suffice to show that $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c + \sum_{s<t} ||g_s||^2 + c' ||\nabla F(w_t)||^2}}$$

L-smoothness $\Longrightarrow$ $\mathbb{E}[||g_t||^2] = poly(t)$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - How to obtain a convergence rate from the following descent inequality?
    $$\mathbb{E}[\textstyle\sum_{t \le T} \tilde{\eta}_t ||\nabla F(w_t)||^2] \le F(w_0) - F^* + c \text{ poly} \log(T)$$
  - Would suffice to show that $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$

Stochastic Optimization with Adaptive Step Sizes – Key Ideas

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - How to obtain a convergence rate from the following descent inequality?
    $$\mathbb{E}[\sum_{t\le T} \tilde{\eta}_t ||\nabla F(w_t)||^2] \le F(w_0) - F^* + c \text{ poly} \log(T)$$
  - Would suffice to show that $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c + \sum_{s<t} ||g_s||^2 + c' ||\nabla F(w_t)||^2}}$$

L-smoothness $\Longrightarrow$ $\mathbb{E}[||g_t||^2] = poly(t)$

Bound may be *exponentially worse* under (L_0, L_1)-smoothness

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c+\sum_{s<t}||g_s||^2+c'||\nabla F(w_t)||^2}}$$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - How to obtain a convergence rate from the following descent inequality?

$$\mathbb{E}[\sum_{t\leq T}\tilde{\eta}_t||\nabla F(w_t)||^2] \leq F(w_0) - F^* + c \text{ poly} \log(T)$$

  - Would suffice to show that $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$

An idea: **Suppose** $\tilde{\eta}_T$ and $\{||\nabla F(w_t)||^2\}_{\{t\leq T\}}$ were **positively correlated**…

Stochastic Optimization with Adaptive Step Sizes – Key Ideas

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c + \sum_{s<t} ||g_s||^2 + c' ||\nabla F(w_t)||^2}}$$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - How to obtain a convergence rate from the following descent inequality?
    $$\mathbb{E}[\textstyle\sum_{t \le T} \tilde{\eta}_t ||\nabla F(w_t)||^2] \le F(w_0) - F^* + c \text{ poly} \log(T)$$
  - Would suffice to show that $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$

An idea: **Suppose** $\tilde{\eta}_T$ and $\{||\nabla F(w_t)||^2\}_{\{t \le T\}}$ were **positively correlated**…

If this were true***…***

$$poly \log(T) \ge \mathbb{E}\left[\sum_{t \le T} \tilde{\eta}_t ||\nabla F(w_t)||^2\right] \gtrsim \mathbb{E}[\tilde{\eta}_T]\mathbb{E}\left[\sum_{t \le T} ||\nabla F(w_t)||^2\right] \gtrsim \frac{\mathbb{E}[\sum_{t \le T} ||\nabla F(w_t)||^2]}{\sqrt{b_0^2 + T\sigma_0^2 + (1 + \sigma_1^2)\mathbb{E}[\sum_{t \le T} ||\nabla F(w_t)||^2]}}$$

Descent inequality

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c + \sum_{s<t} ||g_s||^2 + c' ||\nabla F(w_t)||^2}}$$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - How to obtain a convergence rate from the following descent inequality?
    $$\mathbb{E}[\sum_{t \le T} \tilde{\eta}_t ||\nabla F(w_t)||^2] \le F(w_0) - F^* + c \text{ poly} \log(T)$$
  - Would suffice to show that $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$

An idea: **Suppose** $\tilde{\eta}_T$ and $\{||\nabla F(w_t)||^2\}_{\{t \le T\}}$ were **positively correlated**…

If this were true***…***

$$poly \log(T) \ge \mathbb{E}\left[\sum_{t \le T} \tilde{\eta}_t ||\nabla F(w_t)||^2\right] \gtrsim \mathbb{E}[\tilde{\eta}_T]\mathbb{E}\left[\sum_{t \le T} ||\nabla F(w_t)||^2\right] \gtrsim \frac{\mathbb{E}[\sum_{t \le T} ||\nabla F(w_t)||^2]}{\sqrt{b_0^2 + T\sigma_0^2 + (1 + \sigma_1^2)\mathbb{E}[\sum_{t \le T} ||\nabla F(w_t)||^2]}}$$

Positive correlation + (roughly) decreasing η_t

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c+\sum_{s<t}||g_s||^2+c'||\nabla F(w_t)||^2}}$$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - How to obtain a convergence rate from the following descent inequality?
    $$\mathbb{E}[\textstyle\sum_{t\leq T}\tilde{\eta}_t||\nabla F(w_t)||^2] \leq F(w_0) - F^* + c\text{ poly}\log(T)$$
  - Would suffice to show that $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$

An idea: **Suppose** $\tilde{\eta}_T$ and $\{||\nabla F(w_t)||^2\}_{\{t\leq T\}}$ were **positively correlated**…

If this were true***…***

$$poly\log(T) \geq \mathbb{E}\left[\sum_{t\leq T}\tilde{\eta}_t||\nabla F(w_t)||^2\right] \gtrsim \mathbb{E}[\tilde{\eta}_T]\mathbb{E}\left[\sum_{t\leq T}||\nabla F(w_t)||^2\right] \gtrsim \frac{\mathbb{E}[\sum_{t\leq T}||\nabla F(w_t)||^2]}{\sqrt{b_0^2+T\sigma_0^2+(1+\sigma_1^2)\mathbb{E}[\sum_{t\leq T}||\nabla F(w_t)||^2]}}$$

Jensen's

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c+\sum_{s<t}||g_s||^2+c'||\nabla F(w_t)||^2}}$$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - How to obtain a convergence rate from the following descent inequality?
$$\mathbb{E}[\textstyle\sum_{t\le T}\tilde{\eta}_t||\nabla F(w_t)||^2] \le F(w_0) - F^* + c \text{ poly} \log(T)$$
  - Would suffice to show that $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$

An idea: **Suppose** $\tilde{\eta}_T$ and $\{||\nabla F(w_t)||^2\}_{\{t\le T\}}$ were **positively correlated**…

If this were true…

$$poly \log(T) \ge \mathbb{E}\left[\sum_{t\le T}\tilde{\eta}_t||\nabla F(w_t)||^2\right] \gtrsim \frac{\mathbb{E}[\sum_{t\le T}||\nabla F(w_t)||^2]}{\sqrt{b_0^2 + T\sigma_0^2 + (1+\sigma_1^2)\mathbb{E}[\sum_{t\le T}||\nabla F(w_t)||^2]}}$$

A **quadratic inequality** in $\sqrt{\mathbb{E}[\sum_{t\le T}||\nabla F(w_t)||^2]}$! $\Longrightarrow \mathbb{E}[\sum_{t\le T}||\nabla F(w_t)||^2] = \tilde{O}(\sqrt{T})$

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c+\sum_{s<t}||g_s||^2+c'||\nabla F(w_t)||^2}}$$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - How to obtain a convergence rate from the following descent inequality?
    $$\mathbb{E}[\textstyle\sum_{t\le T}\tilde{\eta}_t||\nabla F(w_t)||^2] \le F(w_0) - F^* + c \text{ poly} \log(T)$$
  - Would suffice to show that $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$

An idea: **Suppose** $\tilde{\eta}_T$ and $\{||\nabla F(w_t)||^2\}_{\{t\le T\}}$ were **positively correlated**…

If this were true***…***

$$poly\log(T) \ge \mathbb{E}\left[\sum_{t\le T}\tilde{\eta}_t||\nabla F(w_t)||^2\right] \gtrsim \frac{\mathbb{E}[\sum_{t\le T}||\nabla F(w_t)||^2]}{\sqrt{b_0^2 + T\sigma_0^2 + (1+\sigma_1^2)\mathbb{E}[\sum_{t\le T}||\nabla F(w_t)||^2]}}$$

A **quadratic inequality** in $\sqrt{\mathbb{E}[\sum_{t\le T}||\nabla F(w_t)||^2]}$! $\Longrightarrow \mathbb{E}[\sum_{t\le T}||\nabla F(w_t)||^2] = \tilde{O}(\sqrt{T})$

A *stronger* bound than necessary to show $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$…

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c + \sum_{s<t} ||g_s||^2 + c'||\nabla F(w_t)||^2}}$$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - How to obtain a convergence rate from the following descent inequality?
    $$\mathbb{E}[\textstyle\sum_{t \le T} \tilde{\eta}_t ||\nabla F(w_t)||^2] \le F(w_0) - F^* + c \text{ poly} \log(T)$$
  - Would suffice to show that $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$

Problem: increasing $||\nabla F(w_t)||^2$, at least intuitively, could *decrease* $\tilde{\eta}_T$!
$\Longrightarrow$ possibly **negatively** correlated…

If $\tilde{\eta}_T$ and $\{||\nabla F(w_t)||^2\}_{\{t \le T\}}$ were **positively correlated:**

$$poly \log(T) \ge \mathbb{E}\left[\sum_{t \le T} \tilde{\eta}_t ||\nabla F(w_t)||^2\right] \gtrsim \frac{\mathbb{E}[\sum_{t \le T} ||\nabla F(w_t)||^2]}{\sqrt{b_0^2 + T\sigma_0^2 + (1 + \sigma_1^2)\mathbb{E}[\sum_{t \le T} ||\nabla F(w_t)||^2]}}$$

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c + \sum_{s<t} ||g_s||^2 + c'||\nabla F(w_t)||^2}}$$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - How to obtain a convergence rate from the following descent inequality?
    $$\mathbb{E}[\textstyle\sum_{t<\tau} \tilde{\eta}_t ||\nabla F(w_t)||^2] \leq F(w_0) - F^* + c \text{ poly} \log(T)$$
  - Would suffice to show that $\mathbb{E}[\tilde{\eta}_{\tau-1}] \gtrsim 1/\sqrt{T}$ for some $\mathbb{E}[\tau] = \Omega(T)$

Key Idea 2:
Analyze convergence only until a **stopping time** τ satisfying $\mathbb{E}[\tau] = \Omega(T)$:

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c+\sum_{s<t}||g_s||^2+c'||\nabla F(w_t)||^2}}$$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - How to obtain a convergence rate from the following descent inequality?
  $$\mathbb{E}[\textstyle\sum_{t<\tau}\tilde{\eta}_t||\nabla F(w_t)||^2] \le F(w_0) - F^* + c \text{ poly}\log(T)$$
  - Would suffice to show that $\mathbb{E}[\tilde{\eta}_{\tau-1}] \gtrsim 1/\sqrt{T}$ for some $\mathbb{E}[\tau] = \Omega(T)$

Key Idea 2:
Analyze convergence only until a **stopping time** τ satisfying $\mathbb{E}[\tau] = \Omega(T)$:

$\exists\tau$ w/ $\mathbb{E}[\tau] = \Omega(T)$ such that $\tilde{\eta}_t$ and $\nabla F(w_t)$ are *roughly* positively correlated **before** $\boldsymbol{\tau}$:

$$poly\log(T) \ge \mathbb{E}\left[\sum_{t<\tau}\tilde{\eta}_t||\nabla F(w_t)||^2\right] \gtrsim \frac{\mathbb{E}[\sum_{t<\tau}||\nabla F(w_t)||^2]}{\sqrt{b_0^2 + \frac{T\sigma_0^2 + (1+\sigma_1^2)\mathbb{E}[\sum_{t<\tau}||\nabla F(w_t)||^2]}{\delta}}}$$

Stochastic Optimization with Adaptive Step Sizes – Key Ideas

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c + \sum_{s<t} ||g_s||^2 + c' ||\nabla F(w_t)||^2}}$$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - How to obtain a convergence rate from the following descent inequality?
  $$\mathbb{E}[\textstyle\sum_{t<\tau} \tilde{\eta}_t ||\nabla F(w_t)||^2] \leq F(w_0) - F^* + c \text{ poly} \log(T)$$
  - Would suffice to show that $\mathbb{E}[\tilde{\eta}_{\tau-1}] \gtrsim 1/\sqrt{T}$ for some $\mathbb{E}[\tau] = \Omega(T)$

Key Idea 2:
Analyze convergence only until a **stopping time** τ satisfying $\mathbb{E}[\tau] = \Omega(T)$:

$\exists \tau$ w/ $\mathbb{E}[\tau] = \Omega(T)$ such that $\tilde{\eta}_t$ and $\nabla F(w_t)$ are *roughly* positively correlated **before τ**:

$$poly \log(T) \geq \mathbb{E}\left[\sum_{t<\tau} \tilde{\eta}_t ||\nabla F(w_t)||^2\right] \gtrsim \frac{\mathbb{E}[\sum_{t<\tau} ||\nabla F(w_t)||^2]}{\sqrt{b_0^2 + \frac{T\sigma_0^2 + (1+\sigma_1^2)\mathbb{E}[\sum_{t<\tau} ||\nabla F(w_t)||^2]}{\delta}}}$$

Stopped descent inequality

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c + \sum_{s<t} ||g_s||^2 + c' ||\nabla F(w_t)||^2}}$$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - How to obtain a convergence rate from the following descent inequality?
    $$\mathbb{E}[\textstyle\sum_{t<\tau} \tilde{\eta}_t ||\nabla F(w_t)||^2] \le F(w_0) - F^* + c \text{ poly} \log(T)$$
  - Would suffice to show that $\mathbb{E}[\tilde{\eta}_{\tau-1}] \gtrsim 1/\sqrt{T}$ for some $\mathbb{E}[\tau] = \Omega(T)$

Key Idea 2:
Analyze convergence only until a **stopping time** τ satisfying $\mathbb{E}[\tau] = \Omega(T)$:

$\exists \tau$ w/ $\mathbb{E}[\tau] = \Omega(T)$ such that $\tilde{\eta}_t$ and $\nabla F(w_t)$ are *roughly* positively correlated **before** $\boldsymbol{\tau}$:

$$\text{poly} \log(T) \ge \mathbb{E}\left[\sum_{t<\tau} \tilde{\eta}_t ||\nabla F(w_t)||^2\right] \gtrsim \frac{\mathbb{E}[\sum_{t<\tau} ||\nabla F(w_t)||^2]}{\sqrt{b_0^2 + \frac{T\sigma_0^2 + (1+\sigma_1^2)\mathbb{E}[\sum_{t<\tau} ||\nabla F(w_t)||^2]}{\delta}}}$$

$$\frac{1}{\delta} = \Theta(\mathrm{T})$$

Rough positive correlation

$$\tilde{\eta}_t = \frac{\eta}{\sqrt{c+\sum_{s<t}||g_s||^2+c'||\nabla F(w_t)||^2}}$$

Overcoming the challenges of **adaptive** step sizes

- **Challenge 2**: Step size scaling
  - How to obtain a convergence rate from the following descent inequality?
    $$\mathbb{E}[\textstyle\sum_{t<\tau} \tilde{\eta}_t ||\nabla F(w_t)||^2] \le F(w_0) - F^* + c \text{ poly} \log(T)$$
  - Would suffice to show that $\mathbb{E}[\tilde{\eta}_{\tau-1}] \gtrsim 1/\sqrt{T}$ for some $\mathbb{E}[\tau] = \Omega(T)$

Key Idea 2:
Analyze convergence only until a **stopping time** τ satisfying $\mathbb{E}[\tau] = \Omega(T)$:

$\exists\tau$ such that $\tilde{\eta}_t$ and $\nabla F(w_t)$ are *roughly* positively correlated **before** $\boldsymbol{\tau}$:

$$poly \log(T) \ge \mathbb{E}\left[\sum_{t<\tau} \tilde{\eta}_t ||\nabla F(w_t)||^2\right] \gtrsim \frac{\mathbb{E}[\sum_{t<\tau} ||\nabla F(w_t)||^2]}{\sqrt{b_0^2 + \frac{T\sigma_0^2 + (1+\sigma_1^2)\mathbb{E}[\sum_{t<\tau} ||\nabla F(w_t)||^2]}{\delta}}}$$

$$\frac{1}{\delta} = \Theta(T)$$

Solving $\Longrightarrow \mathbb{E}[\sum_{t<\tau} ||\nabla F(w_t)||^2] = \tilde{O}(T)$
$\Longrightarrow \mathbb{E}[\tilde{\eta}_{\tau-1}] \gtrsim 1/\sqrt{T}$

Convergence under non-uniform smoothness condition

<u>AdaGrad-Norm Algorithm</u>

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t} \|g_t\|^2}} \cdot g_t$$

$\eta \lesssim 1/L_1(1+\sigma_1^2)$

Theorem (COLT'23)

AdaGrad-Norm enjoys a $\min_t \|\nabla F(w_t)\|^2 = \tilde{O}(1/\sqrt{T})$ convergence rate assuming:

- F is (L_0, L_1)-smooth and either:

  $\sigma_1 < 1$ **or** $\sigma_1 \geq 1$ and: *(i)* mini-batch size $\Omega(\sigma_1^2)$, **or**

  (ii) F is "polynomially-bounded"

Non-Convergence of standard algorithms under multiplicative noise

Normalized/Clipped SGD

$$w_{t+1} = w_t - \eta_t g_t \qquad \eta_t = \begin{cases} \frac{\eta}{\gamma + ||g_t||} & \text{Normalized SGD} \\ \frac{\eta}{\max\{\gamma, ||g_t||\}} & \text{Clipped SGD} \end{cases}$$

Theorem (COLT'23)

There is a stochastic gradient oracle which:

- Is *unbiased* and satisfies *affine variance* ($\sigma_0 = 0, \sigma_1 > 1$)
- Yet does not converge with constant probability on a 1-D quadratic function in many parameter regimes
  - E.g., when $\gamma = 0$, diverges for *any* choice of η

Key Takeaway

- AdaGrad-Norm works in settings where many standard algorithms for (L_0, L_1)-optimization can fail!

Concurrent work at COLT'23

Concurrent work in COLT'23 [Wang-Zhang-Ma-Chen'23]

- Analyze AdaGrad under (L_0, L_1)-smoothness and affine variance
- Establish convergence without some technical restrictions needed for our analysis
- They bound the bias between g_t and η_t using an auxiliary function which *telescopes*

Gives descent inequality over *entire* time horizon $[T]$

- We give a different analysis relying on a carefully-constructed stopping time τ

"Decorrelates" gradients from steps-sizes before τ

Useful in settings where descent inequality holds only over a *random* $S \subset [T]$

Convergence under non-uniform smoothness condition

<u>AdaGrad-Norm Algorithm</u>

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{b_0^2 + \sum_{s=1}^{t} \|g_t\|^2}} \cdot g_t$$

$\eta \lesssim 1/L_1(1 + \sigma_1^2)$

Theorem (COLT'23)

AdaGrad-Norm enjoys a $\min_t \|\nabla F(w_t)\|^2 = \tilde{O}(1/\sqrt{T})$ convergence rate assuming:

- F is (L_0, L_1)-smooth and either:

  $\sigma_1 < 1$ **or** $\sigma_1 \geq 1$ and: *(i)* mini-batch size $\Omega(\sigma_1^2)$, **or**

  (ii) F is "polynomially-bounded"

Thanks for listening!

Any questions can be sent to:

{matthewfaw, litu.rout}@utexas.edu

Beyond Uniform Smoothness: A Stopped Analysis of Adaptive SGD

arXiv:2302.06570