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

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

 $||\nabla F(x) - \nabla F(y)|| \le \frac{L_0}{|x - y||} \quad \forall x, y$

When *F* is twice-differentiable, equivalent to:

 $||\nabla^2 F(x)|| \le \frac{L_0}{V} \,\forall x$

Find a first-order stationary point of a non-convex, (L_0, L_1) -smooth¹ function F:

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

When *F* is twice-differentiable, \approx equivalent to:

 $||\nabla^2 F(x)|| \le \frac{L_0}{L_0} + \frac{L_1}{||\nabla F(x)||} \quad \forall x$

Find a first-order stationary point of a non-convex, (L_0, L_1) -smooth¹ function F:

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

When *F* is twice-differentiable, \approx equivalent to:

 $||\nabla^2 F(x)|| \le \frac{L_0}{L_0} + \frac{L_1}{||\nabla F(x)||} \quad \forall x$

• Standard *L*-smoothness is equivalent to (*L*, 0)-smoothness

Find a first-order stationary point of a non-convex, (L_0, L_1) -smooth¹ function F:

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

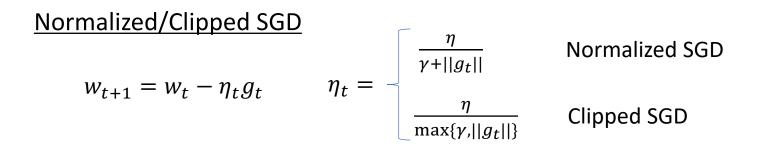
When *F* is twice-differentiable, \approx equivalent to:

 $||\nabla^2 F(x)|| \le \frac{L_0}{L_0} + \frac{L_1}{||\nabla F(x)||} \quad \forall x$

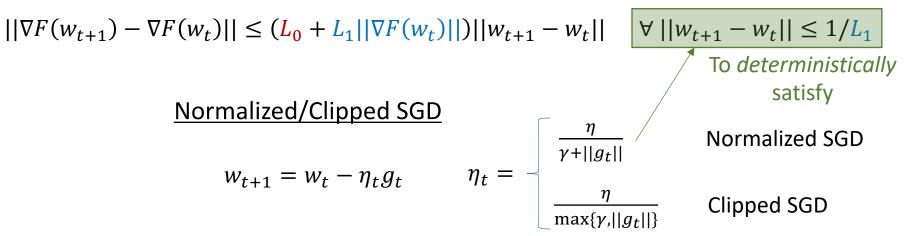
- Standard *L*-smoothness is equivalent to (*L*, 0)-smoothness
- Also captures a wide class of functions which are not uniformly smooth, e.g.:
 - $F(x) = x^c$ for c > 2 (c(c-1), c-1)-smooth
 - $F(x) = e^{c'x}$ for c > 0 (0, c')-smooth

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

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



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



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

Prior work¹ established $\min_{t} \|\nabla F(w_t)\|^2 = O(1/\sqrt{T})$ convergence rate assuming:

•
$$\mathbb{E}[g] = \nabla F(w)$$

• $\sup_{w} ||g - \nabla F(w)||^2 \le \sigma_0^2$

(unbiased stochastic gradient) (bounded *noise support*)

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

$$\begin{split} |\nabla F(w_{t+1}) - \nabla F(w_t)|| &\leq (L_0 + L_1 ||\nabla F(w_t)||) ||w_{t+1} - w_t|| \qquad \forall ||w_{t+1} - w_t|| \leq 1/L_1 \\ & \text{To deterministically satisfy} \\ \\ \frac{\text{Normalized/Clipped SGD}}{w_{t+1} = w_t - \eta_t g_t} \qquad \eta_t = - \begin{bmatrix} \frac{\eta}{\gamma + ||g_t||} & \text{Normalized SGD} \\ \frac{\eta}{\max\{\gamma, ||g_t||\}} & \text{Clipped SGD} \end{bmatrix} \end{split}$$

Prior work² established $\min_{t} \|\nabla F(w_t)\|^2 = O(1/\sqrt{T})$ convergence rate assuming:

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

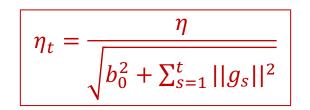
Significantly stronger assumption than is needed in *L*-smooth setting

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

```
||\nabla F(x) - \nabla F(y)|| \le L_0 ||x - y|| \quad \forall x, y
```

AdaGrad-Norm

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

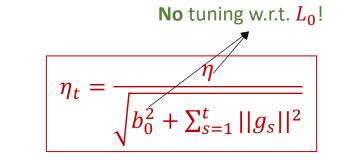


Prior work³ established $\min_{t} \|\nabla F(w_t)\|^2 = O(1/\sqrt{T})$ convergence rate assuming:

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

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

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



AdaGrad-Norm

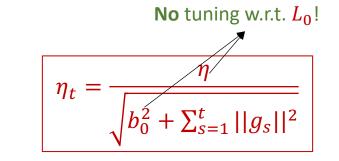
$$w_{t+1} = w_t - \eta_t g_t$$

Prior work³ established $\min_{t} \|\nabla F(w_t)\|^2 = O(1/\sqrt{T})$ convergence rate assuming:

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

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

$$||\nabla F(x) - \nabla F(y)|| \le \frac{L_0}{|x - y||} \quad \forall x, y$$



AdaGrad-Norm

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

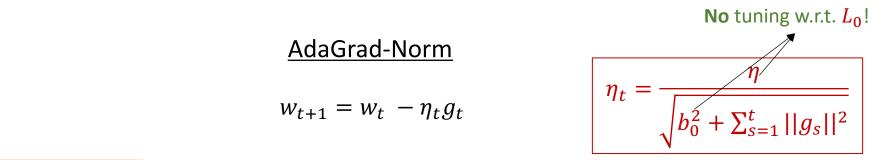
Prior work³ established $\min_{t} \|\nabla F(w_t)\|^2 = O(1/\sqrt{T})$ convergence rate assuming:

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

Analysis *heavily* relies on the L_0 -smoothness assumption

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

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



Question

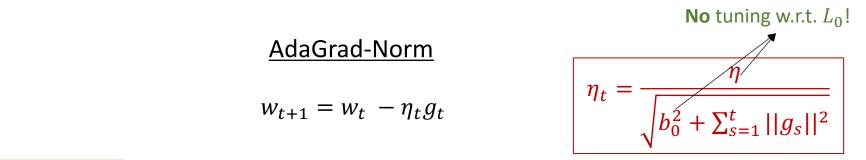
Given that AdaGrad-Norm adapts to the smoothness parameter L_0 automatically...

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

- (L_0, L_1) -smoothness
- Affine variance

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

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



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

les!

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

- **Challenge 1**: Bias + affine variance
 - Step size η_t depends on past and current stochastic gradients.
 - \Rightarrow Obtaining a useful descent lemma from smoothness becomes challenging $\eta_t \|\nabla F(w_t)\|^2 \leq F(w_t) - F(w_{t+1}) + \frac{\eta_t \langle \nabla F(w_t), \nabla F(w_t) - g_t \rangle}{2} + \frac{(L_0 + L_1 ||\nabla F(w_t)||) \eta_t^2}{2} \|g_t\|^2$ Not mean-zero!

- **Challenge 1**: Bias + affine variance
 - Step size η_t depends on past and current stochastic gradients.
 - \Rightarrow Obtaining a useful descent lemma from smoothness becomes challenging $\eta_t \|\nabla F(w_t)\|^2 \le F(w_t) - F(w_{t+1}) + \eta_t \langle \nabla F(w_t), \nabla F(w_t) - g_t \rangle + \frac{(L_0 + L_1 ||\nabla F(w_t)||)\eta_t^2}{2} \|g_t\|^2$

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

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

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

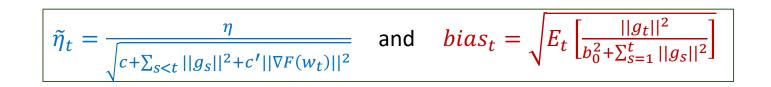
 $\|\tilde{\eta}_t(1 - \sigma_1 \cdot bias_t)\|\nabla F(w_t)\|^2 \le \mathbb{E}_t[F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t[\eta_t^2 \|g_t\|^2]$

Step-size "proxy" Lower bound for $\mathbb{E}_t[\eta_t]$

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

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

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

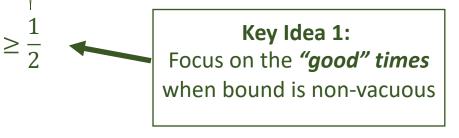


Descent direction $-\eta_t g_t$ is **biased!**

19

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

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

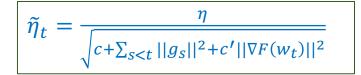


$$\tilde{\eta}_{t} = \frac{\eta}{\sqrt{c + \sum_{s < t} ||g_{s}||^{2} + c'||\nabla F(w_{t})||^{2}}} \quad \text{and} \quad bias_{t} = \sqrt{E_{t} \left[\frac{||g_{t}||^{2}}{b_{0}^{2} + \sum_{s=1}^{t} ||g_{s}||^{2}}\right]}$$

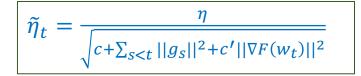
- **Challenge 1**: Bias + affine variance
 - Step size η_t depends on past and current stochastic gradients.
 - Especially challenging under **affine variance** $\tilde{\eta}_t (1 - \sigma_1 \cdot bias_t) \|\nabla F(w_t)\|^2 \leq \mathbb{E}_t [F(w_t) - F(w_{t+1})] + c \cdot \mathbb{E}_t [\eta_t^2 \|g_t\|^2]$
 - Most times are (typically) "good" ⇒ descent inequality (roughly) of the form:

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

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

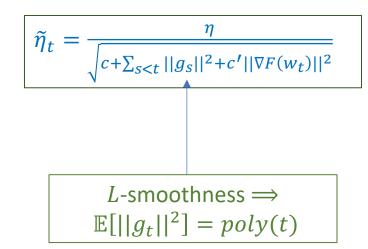


- Challenge 2: Step size scaling
 - How to obtain a convergence rate from the following descent inequality? $\mathbb{E}[\sum_{t \leq T} \tilde{\eta}_t ||\nabla F(w_t)||^2] \leq F(w_0) - F^* + c \text{ poly } \log(T)$

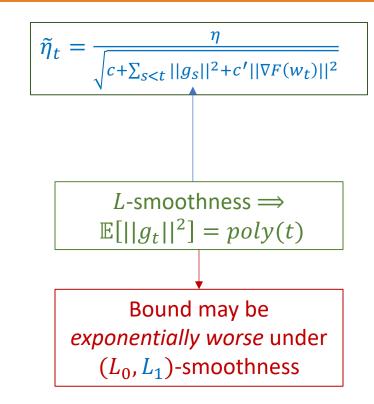


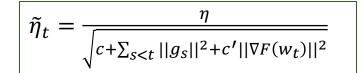
- Challenge 2: Step size scaling
 - How to obtain a convergence rate from the following descent inequality? $\mathbb{E}[\sum_{t \leq T} \tilde{\eta}_t ||\nabla F(w_t)||^2] \leq F(w_0) - F^* + c \text{ poly } \log(T)$
 - Would suffice to show that $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$

- 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 $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$

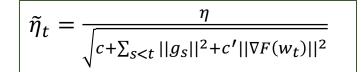


- Challenge 2: Step size scaling
 - How to obtain a convergence rate from the following descent inequality? $\mathbb{E}[\sum_{t \leq T} \tilde{\eta}_t ||\nabla F(w_t)||^2] \leq F(w_0) - F^* + c \text{ poly } \log(T)$
 - Would suffice to show that $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$





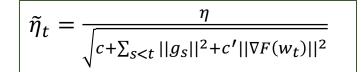
- Challenge 2: Step size scaling
 - How to obtain a convergence rate from the following descent inequality? $\mathbb{E}[\sum_{t \leq T} \tilde{\eta}_t ||\nabla F(w_t)||^2] \leq F(w_0) - F^* + c \text{ poly } \log(T)$
 - Would suffice to show that $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$



- 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 $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$

$$If \text{ this were true}...$$

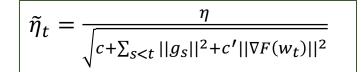
$$poly \log(T) \ge \mathbb{E}\left[\sum_{t \le T} \tilde{\eta}_t ||\nabla F(w_t)||^2\right] \gtrsim \mathbb{E}[\tilde{\eta}_T] \mathbb{E}\left[\sum_{t \le T} ||\nabla F(w_t)||^2\right] \gtrsim \frac{\mathbb{E}[\sum_{t \le T} ||\nabla F(w_t)||^2]}{\sqrt{b_0^2 + T\sigma_0^2 + (1 + \sigma_1^2)\mathbb{E}[\sum_{t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



- 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 $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$

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

$$\begin{aligned} \text{Positive correlation} + (\text{roughly}) \\ \text{decreasing } \eta_t \end{aligned}$$

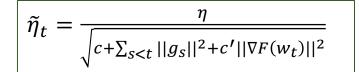


- Challenge 2: Step size scaling
 - How to obtain a convergence rate from the following descent inequality? $\mathbb{E}[\sum_{t \leq T} \tilde{\eta}_t ||\nabla F(w_t)||^2] \leq F(w_0) - F^* + c \text{ poly } \log(T)$
 - Would suffice to show that $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$

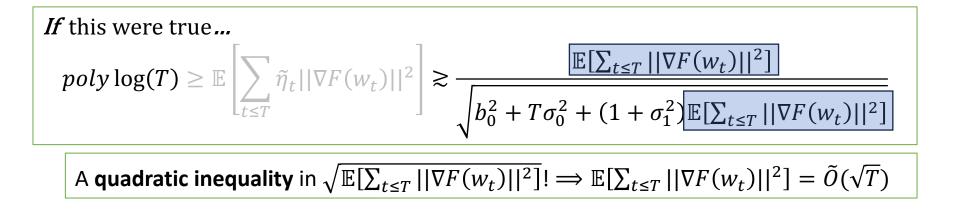
$$If \text{ this were true}...$$

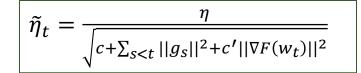
$$poly \log(T) \ge \mathbb{E}\left[\sum_{t \le T} \tilde{\eta}_t ||\nabla F(w_t)||^2\right] \gtrsim \mathbb{E}[\tilde{\eta}_T] \mathbb{E}\left[\sum_{t \le T} ||\nabla F(w_t)||^2\right] \gtrsim \frac{\mathbb{E}[\sum_{t \le T} ||\nabla F(w_t)||^2]}{\sqrt{b_0^2 + T\sigma_0^2 + (1 + \sigma_1^2)\mathbb{E}[\sum_{t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

$$Jensen's$$



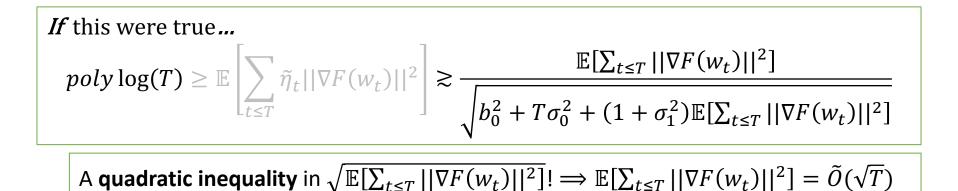
- Challenge 2: Step size scaling
 - How to obtain a convergence rate from the following descent inequality? $\mathbb{E}[\sum_{t \leq T} \tilde{\eta}_t ||\nabla F(w_t)||^2] \leq F(w_0) - F^* + c \text{ poly } \log(T)$
 - Would suffice to show that $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$





- 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...



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

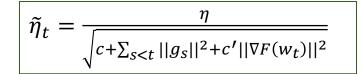
- Challenge 2: Step size scaling
 - How to obtain a convergence rate from the following descent inequality? $\mathbb{E}[\sum_{t \leq T} \tilde{\eta}_t ||\nabla F(w_t)||^2] \leq F(w_0) - F^* + c \text{ poly } \log(T)$
 - Would suffice to show that $\mathbb{E}[\tilde{\eta}_T] \gtrsim 1/\sqrt{T}$

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

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

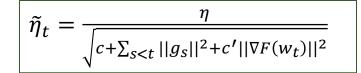
 $\tilde{\eta}_t =$

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



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



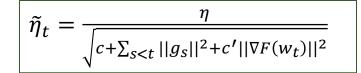
- Challenge 2: Step size scaling
 - How to obtain a convergence rate from the following descent inequality? $\mathbb{E}[\sum_{t < \tau} \tilde{\eta}_t ||\nabla F(w_t)||^2] \le F(w_0) - F^* + c \text{ poly } \log(T)$
 - Would suffice to show that $\mathbb{E}[\tilde{\eta}_{\tau-1}] \gtrsim 1/\sqrt{T}$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

Key Idea 2:

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

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

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



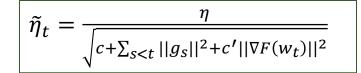
- Challenge 2: Step size scaling
 - How to obtain a convergence rate from the following descent inequality? $\mathbb{E}[\sum_{t < \tau} \tilde{\eta}_t ||\nabla F(w_t)||^2] \le F(w_0) - F^* + c \text{ poly } \log(T)$
 - Would suffice to show that $\mathbb{E}[\tilde{\eta}_{\tau-1}] \gtrsim 1/\sqrt{T}$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

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

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

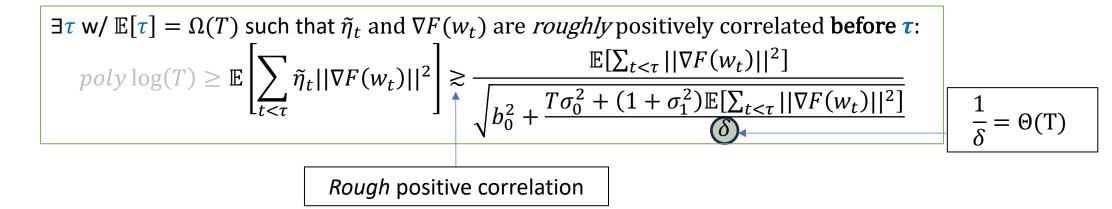
$$poly \log(T) \ge \mathbb{E}\left[\sum_{t < \tau} \tilde{\eta}_t ||\nabla F(w_t)||^2\right] \gtrsim \frac{\mathbb{E}[\sum_{t < \tau} ||\nabla F(w_t)||^2]}{\sqrt{b_0^2 + \frac{T\sigma_0^2 + (1 + \sigma_1^2)\mathbb{E}[\sum_{t < \tau} ||\nabla F(w_t)||^2]}}{\delta}$$

Stopped descent inequality



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



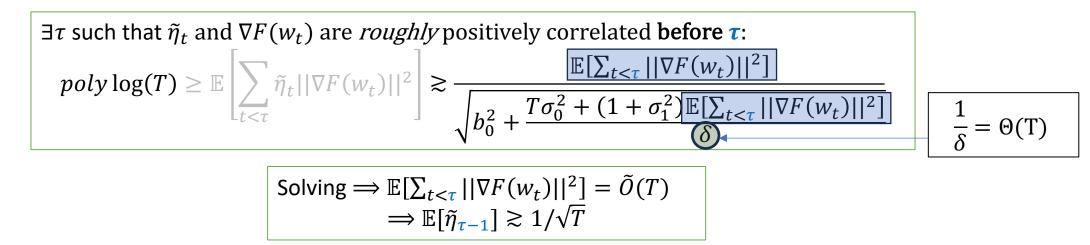


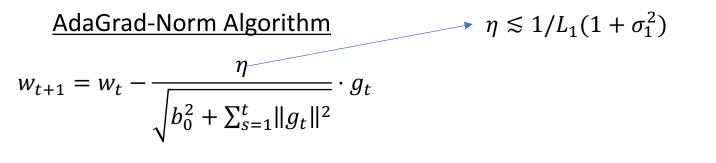
$\tilde{\eta}_t = \frac{\eta}{\sqrt{c + \sum_{s < t} ||g_s||^2 + c'||\nabla F(w_t)||^2}}$

Overcoming the challenges of **adaptive** step sizes

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



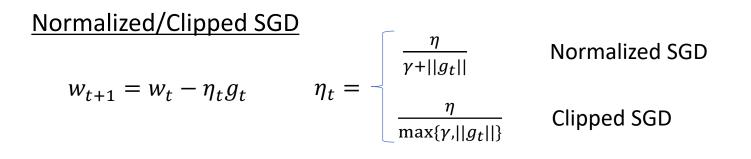


Theorem (COLT'23)

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 $\sigma_1 \ge 1$ and: (i) mini-batch size $\Omega(\sigma_1^2)$, or (ii) *F* is "polynomially-bounded"



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

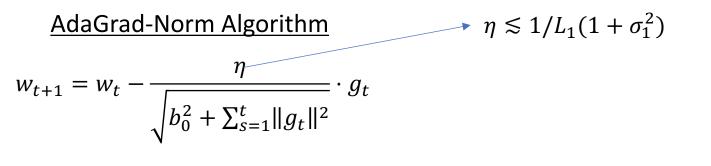
- Analyze AdaGrad under (L_0, L_1) -smoothness and affine variance
- Establish convergence without some technical restrictions needed for our analysis
- They bound the bias between g_t and η_t using an auxiliary function which *telescopes*

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

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

"Decorrelates" gradients from steps-sizes before au

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



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 $\sigma_1 \ge 1$ 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

