INFORMS OPTIMIZATION SOCIETY CONFERENCE 2026 · ATLANTA, GA

Stochastic Auto-Conditioned Fast Gradient Methods with Optimal Rates

Yao Ji & Guanghui (George) Lan
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Tech

PART I

Problem Setup

What This Talk Is About

We study composite stochastic optimization:

$$\Phi^* = \min_{x \in X} \{f(x) + h(x)\}, \qquad X \subseteq \mathbb{R}^n \text{ closed convex}$$

- f: convex, differentiable, L-smooth with unknown L
- h: closed, convex, prox_h efficiently computable
- Gradients via stochastic oracle $G(x, \xi)$

Main Question

Can we achieve optimal convergence rates without knowing L and without line search?

Accelerated Unbounded Stochastic

	KREISLER ET AL., '24	OURS
Distribution	Restricted: subgaussian (σ_x)	Bounded variance

Accelerated Unbounded Stochastic

	KREISLER ET AL., '24	OURS
Distribution	Restricted: subgaussian (σ_x)	Bounded variance
Rate	Sub-optimal $\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{\varepsilon}} + \frac{\sigma^2}{\varepsilon^2} + \frac{\sigma_*^2}{\varepsilon}\right)$	Optimal $\mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} + \frac{\sigma^2}{\varepsilon^2}\right)$

Accelerated Unbounded Stochastic

	KREISLER ET AL., '24	OURS
Distribution	Restricted: subgaussian (σ_x)	Bounded variance
Rate	Sub-optimal $\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{\varepsilon}} + \frac{\sigma^2}{\varepsilon^2} + \frac{\sigma_*^2}{\varepsilon}\right)$	Optimal $\mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} + \frac{\sigma^2}{\varepsilon^2}\right)$
Parameter	Need subgaussian $\sigma(x)$	Variance estimation

Accelerated Unbounded Stochastic

	KREISLER ET AL., '24	OURS
Distribution	Restricted: subgaussian (σ_x)	Bounded variance
Rate	Sub-optimal $\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{\varepsilon}} + \frac{\sigma^2}{\varepsilon^2} + \frac{\sigma_*^2}{\varepsilon}\right)$	Optimal $\mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} + \frac{\sigma^2}{\varepsilon^2}\right)$
Parameter	Need subgaussian $\sigma(x)$	Variance estimation
Gradient	Bounded gradient, minibatch	No such assumption

Accelerated Unbounded Stochastic

	KREISLER ET AL., '24	OURS
Distribution	Restricted: subgaussian (σ_x)	Bounded variance
Rate	Sub-optimal $\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{\varepsilon}} + \frac{\sigma^2}{\varepsilon^2} + \frac{\sigma_*^2}{\varepsilon}\right)$	Optimal $\mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}} + \frac{\sigma^2}{\varepsilon^2}\right)$
Parameter	Need subgaussian $\sigma(x)$	Variance estimation
Gradient	Bounded gradient, minibatch	No such assumption
Horizon	Need total iteration N	N free, auto-conditioned

PART II

The Algorithm

Recap: Auto-Conditioned Fast Gradient Method (Li & Lan, 2024)

$$z_k = \operatorname*{argmin}_{z \in X} \left\{ \langle \nabla f(x_{k-1}), z \rangle + \frac{1}{2\eta_k} \|z - y_{k-1}\|^2 \right\}$$

$$x_k = \frac{\tau_k}{1+\tau_k} x_{k-1} + \frac{1}{1+\tau_k} z_k \qquad y_k = (1-\beta_k) y_{k-1} + \beta_k z_k$$

where

$$\eta_k \propto \min \left\{ \frac{k\eta_{k-1}}{k-1}, \frac{k}{\bar{L}_{k-1}} \right\}$$

$$\bar{L}_{k-1} = \frac{\|\nabla f(x_{k-1}) - \nabla f(x_{k-2})\|^2}{2\left[f(x_{k-2}) - f(x_{k-1}) - \langle \nabla f(x_{k-1}), x_{k-2} - x_{k-1} \rangle\right]}$$

Stochastic Auto-Conditioned Fast Gradient Method

$$z_k = \operatorname*{argmin}_{z \in X} \left\{ \langle G(x_{k-1}, \xi), z \rangle + \frac{1}{2\eta_k} \|z - y_{k-1}\|^2 \right\}$$

$$x_k = \frac{\tau_k}{1+\tau_k} x_{k-1} + \frac{1}{1+\tau_k} z_k \qquad y_k = (1-\beta_k) y_{k-1} + \beta_k z_k$$

where

$$\eta_k \propto \min \left\{ \frac{k\eta_{k-1}}{k-1}, \frac{k}{\bar{L}_{k-1}} \right\}$$

$$\bar{L}_{k-1} = \frac{\left\| G(x_{k-1}, \bar{\xi}) - G(x_{k-2}, \bar{\xi}) \right\|^2}{2\left[F(x_{k-2}, \bar{\xi}) - F(x_{k-1}, \bar{\xi}) - \left\langle G(x_{k-1}, \bar{\xi}), x_{k-2} - x_{k-1} \right\rangle \right]}$$

Stochastic Auto-Conditioned Fast Gradient Method

$$z_k = \operatorname*{argmin}_{z \in X} \left\{ \langle G(x_{k-1}, \xi), z \rangle + \frac{1}{2\eta_k} \|z - y_{k-1}\|^2 \right\}$$

$$x_k = \frac{\tau_k}{1+\tau_k} x_{k-1} + \frac{1}{1+\tau_k} z_k \qquad y_k = (1-\beta_k) y_{k-1} + \beta_k z_k$$

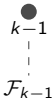
ATTEMPT FAILS!

where

$$\eta_k \propto \min \left\{ \frac{k\eta_{k-1}}{k-1}, \frac{k}{\bar{L}_{k-1}} \right\}$$

$$\bar{L}_{k-1} = \frac{\left\| G(x_{k-1}, \bar{\xi}) - G(x_{k-2}, \bar{\xi}) \right\|^2}{2\left[F(x_{k-2}, \bar{\xi}) - F(x_{k-1}, \bar{\xi}) - \left\langle G(x_{k-1}, \bar{\xi}), x_{k-2} - x_{k-1} \right\rangle \right]}$$

Three Batch Types & Filtration

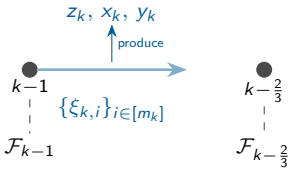


Three Batch Types & Filtration

Gradient Est.

Samples $\{\xi_{k,i}\}_{i=1}^{m_k}$

Compute G_k z_k, x_k, y_k



Type I: iterate update batch $\xi_{k,i}$

Gradient estimate

$$G_k = \frac{1}{m_k} \sum_{i=1}^{m_k} G(x_{k-1}, \xi_{k,i})$$

Type I: iterate update batch $\xi_{k,i}$

Gradient estimate

$$G_k = \frac{1}{m_k} \sum_{i=1}^{m_k} G(x_{k-1}, \xi_{k,i})$$

Proximal step

$$z_k = \underset{z \in X}{\operatorname{argmin}} \left\{ \langle G_k, z \rangle + h(z) + \frac{1}{2\eta_k} \|y_{k-1} - z\|^2 \right\}$$

Type I: iterate update batch $\xi_{k,i}$

Gradient estimate

$$G_k = \frac{1}{m_k} \sum_{i=1}^{m_k} G(x_{k-1}, \xi_{k,i})$$

Proximal step

$$z_k = \operatorname*{argmin}_{z \in X} \left\{ \langle G_k, z \rangle + h(z) + \frac{1}{2\eta_k} \|y_{k-1} - z\|^2 \right\}$$

Output solution

$$x_k = \frac{z_k + \tau_k x_{k-1}}{1 + \tau_k}, \qquad \tau_k = \frac{k}{2}$$

Type I: iterate update batch $\xi_{k,i}$

Gradient estimate

$$G_k = \frac{1}{m_k}\sum_{i=1}^{m_k} G(x_{k-1}, \xi_{k,i})$$

Proximal step

$$z_k = \operatorname*{argmin}_{z\in X}\left\{\langle G_k, z\rangle + h(z) + \frac{1}{2\eta_k}\|y_{k-1}-z\|^2\right\}$$

Output solution

$$x_k = \frac{z_k + \tau_k x_{k-1}}{1+\tau_k}, \qquad \tau_k = \frac{k}{2}$$

Output center

$$y_k = (1-\beta_k)y_{k-1} + \beta_k z_k, \quad \beta_1 = 0,\ \beta_k \equiv \beta$$

Type I: iterate update batch $\xi_{k,i}$

Gradient estimate

$$G_k = \frac{1}{m_k}\sum_{i=1}^{m_k} G(x_{k-1}, \xi_{k,i})$$

Proximal step

$$z_k = \operatorname*{argmin}_{z\in X}\left\{\langle G_k, z\rangle + h(z) + \frac{1}{2\eta_k}\|y_{k-1}-z\|^2\right\}$$

Output solution

$$x_k = \frac{z_k + \tau_k x_{k-1}}{1+\tau_k}, \qquad \tau_k = \frac{k}{2}$$

Output center

$$y_k = (1-\beta_k)y_{k-1} + \beta_k z_k, \quad \beta_1 = 0,\ \beta_k \equiv \beta$$

For the next iteration: compute $\eta_{k+1}, m_{k+1}, n_{k+1}$

Three Batch Types & Filtration

Gradient Est.

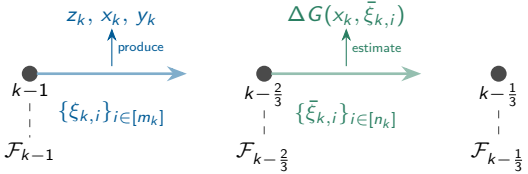
Samples $\{\xi_{k,i}\}_{i=1}^{m_k}$

Compute G_k z_k, x_k, y_k

Stepsize update. I

Samples $\{\bar{\xi}_{k,i}\}$

Compute ΔG



Three Batch Types & Filtration

Gradient Est.

Samples $\{\xi_{k,i}\}_{i=1}^{m_k}$

Compute G_k z_k, x_k, y_k

Stepsize update. I

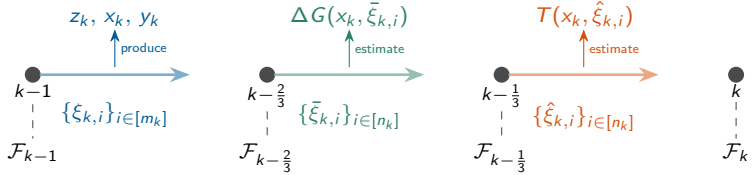
Samples $\{\bar{\xi}_{k,i}\}$

Compute ΔG

Stepsize update. II

Samples $\{\hat{\xi}_{k,i}\}$

Compute T



Type II: stepsize η_{k+1} update batches $\bar{\xi}_{k,i}$, $\widehat{\xi}_{k,i}$

Type II: stepsize η_{k+1} update batches $\bar{\xi}_{k,i}$, $\widehat{\xi}_{k,i}$

1. Stochastic gradient difference:

$$\Delta G(x_k, \bar{\xi}_k) := \frac{1}{n_k} \sum_{i=1}^{n_k} \left[G(x_k, \bar{\xi}_{k,i}) - G(x_{k-1}, \bar{\xi}_{k,i}) \right].$$

Type II: stepsize η_{k+1} update batches $\bar{\xi}_{k,i}$, $\widehat{\xi}_{k,i}$

1. Stochastic gradient difference:

$$\Delta G(x_k, \bar{\xi}_k) := \tfrac{1}{n_k} \textstyle\sum_{i=1}^{n_k} \left[G(x_k, \bar{\xi}_{k,i}) - G(x_{k-1}, \bar{\xi}_{k,i}) \right].$$

2. Empirical first-orde Taylor remainder:

$$T(x_k, \widehat{\xi}_k) := \frac{1}{n_k} \sum_{i=1}^{n_k} \left[F(x_{k-1}, \widehat{\xi}_{k,i}) - F(x_k, \widehat{\xi}_{k,i}) - \langle G(x_k, \widehat{\xi}_{k,i}), x_{k-1} - x_k \rangle \right]$$

Type II: stepsize η_{k+1} update batches $\bar{\xi}_{k,i}$, $\widehat{\xi}_{k,i}$

1. Stochastic gradient difference:

$$\Delta G(x_k, \bar{\xi}_k) := \frac{1}{n_k} \sum_{i=1}^{n_k} \left[G(x_k, \bar{\xi}_{k,i}) - G(x_{k-1}, \bar{\xi}_{k,i}) \right].$$

2. Empirical first-orde Taylor remainder:

$$T(x_k, \widehat{\xi}_k) := \frac{1}{n_k} \sum_{i=1}^{n_k} \left[F(x_{k-1}, \widehat{\xi}_{k,i}) - F(x_k, \widehat{\xi}_{k,i}) - \langle G(x_k, \widehat{\xi}_{k,i}), x_{k-1} - x_k \rangle \right]$$

3. The ratio: empirical local cocoercivity-based smoothness estimator

$$\bar{L}_k := \begin{cases} \frac{\|\Delta G(x_k, \bar{\xi}_{k,i})\|^2}{2T(x_k, \widehat{\xi}_{k,i})}, & \text{if } T(x_k, \widehat{\xi}_{k,i}) > 0, \\ 0, & \text{if } T(x_k, \widehat{\xi}_{k,i}) = 0. \end{cases}$$

Type II: stepsize η_{k+1} update batches $\bar{\xi}_{k,i}$, $\widehat{\xi}_{k,i}$

1. Stochastic gradient difference:

$$\Delta G(x_k, \bar{\xi}_k) := \tfrac{1}{n_k} \textstyle\sum_{i=1}^{n_k} \left[G(x_k, \bar{\xi}_{k,i}) - G(x_{k-1}, \bar{\xi}_{k,i}) \right].$$

2. Empirical first-orde Taylor remainder:

$$T(x_k, \widehat{\xi}_k) := \frac{1}{n_k} \sum_{i=1}^{n_k} \left[F(x_{k-1}, \widehat{\xi}_{k,i}) - F(x_k, \widehat{\xi}_{k,i}) - \langle G(x_k, \widehat{\xi}_{k,i}), x_{k-1} - x_k \rangle \right]$$

3. The ratio: empirical local cocoercivity-based smoothness estimator

$$\bar{L}_k := \begin{cases} \frac{\|\Delta G(x_k, \bar{\xi}_{k,i})\|^2}{2T(x_k, \widehat{\xi}_{k,i})}, & \text{if } T(x_k, \widehat{\xi}_{k,i}) > 0, \\ 0, & \text{if } T(x_k, \widehat{\xi}_{k,i}) = 0. \end{cases}$$

> Computes the ratio of **stochastic gradient variation** to **stochastic local curvature**.

Type II: stepsize η_{k+1} update batches $\bar{\xi}_{k,i}$, $\widehat{\xi}_{k,i}$

1. Stochastic gradient difference:

$$\Delta G(x_k, \bar{\xi}_k) := \frac{1}{n_k} \sum_{i=1}^{n_k} \left[G(x_k, \bar{\xi}_{k,i}) - G(x_{k-1}, \bar{\xi}_{k,i}) \right].$$

2. Empirical first-orde Taylor remainder:

$$T(x_k, \widehat{\xi}_k) := \frac{1}{n_k} \sum_{i=1}^{n_k} \left[F(x_{k-1}, \widehat{\xi}_{k,i}) - F(x_k, \widehat{\xi}_{k,i}) - \langle G(x_k, \widehat{\xi}_{k,i}), x_{k-1} - x_k \rangle \right]$$

3. The ratio: empirical local cocoercivity-based smoothness estimator

$$\bar{L}_k := \begin{cases} \frac{\|\Delta G(x_k, \bar{\xi}_{k,i})\|^2}{2T(x_k, \widehat{\xi}_{k,i})}, & \text{if } T(x_k, \widehat{\xi}_{k,i}) > 0, \\ 0, & \text{if } T(x_k, \widehat{\xi}_{k,i}) = 0. \end{cases}$$

- Computes the ratio of **stochastic gradient variation** to **stochastic local curvature**.
- $\bar{L}_k$ cleanly proxies the local Lipschitz constant near x_k to set the next stepsize η_{k+1}.

Stochastic adaptive stepsize η_{k+1} update

Using $\bar{L}_k$, we define the stepsize recursively as follows. Let $\eta_1 > 0$ and define

$$\eta_2 = \min\left\{\frac{1}{16\beta\bar{L}_1}, 2(1-\beta)\eta_1\right\}, \quad \eta_{k+1} = \min\left\{\frac{k}{16\bar{L}_k}, \frac{(k+1)\eta_k}{k}\right\}, \quad \forall\, k \geq 2, \text{ a.s.}$$

- Large $\bar{L}_k \Rightarrow$ small η_{k+1}: conservative step near curved regions x_k
- Small $\bar{L}_k \Rightarrow \eta_{k+1}$ may grow as $\mathcal{O}(k)$: accelerated regime
- No knowledge of global L at any point in the algorithm

For the next iteration: compute m_{k+1}, n_{k+1}

Adaptive minibatch m_{k+1} for iterate update

Minibatch in parameter free setting

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{v_k^{\max}, \frac{\sigma_k^2}{D_0^2}\right\}\right\rceil\right)$$

Local $\bar{L}_k, \sigma_k^2, v_k^{\max} := \max_{0 \le i \le k} v_i$

- σ_k^2 : local variance of the stochastic gradient error
- $v_k^{\max}$: local variance of the local smoothness estimator

Adaptive minibatch m_{k+1} for iterate update

Minibatch in parameter free setting

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{v_k^{\max}, \frac{\sigma_k^2}{D_0^2}\right\}\right\rceil\right)$$

Local $\bar{L}_k, \sigma_k^2, v_k^{\max} := \max_{0 \le i \le k} v_i$

Minibatch AS-SA Ghadimi & Lan 2016

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{L^2} \cdot \frac{\sigma^2}{D_0^2}\right\rceil\right)$$

Global smoothness L, variance σ^2

- σ_k^2 : local variance of the stochastic gradient error
- $v_k^{\max}$: local variance of the local smoothness estimator

Adaptive minibatch m_{k+1} for iterate update

Minibatch in parameter free setting

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{v_k^{\max}, \frac{\sigma_k^2}{D_0^2}\right\}\right\rceil\right)$$

Local $\bar{L}_k, \sigma_k^2, v_k^{\max} := \max_{0 \le i \le k} v_i$

Minibatch AS-SA Ghadimi & Lan 2016

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{L^2} \cdot \frac{\sigma^2}{D_0^2}\right\rceil\right)$$

Global smoothness L, variance σ^2

- σ_k^2 : local variance of the stochastic gradient error
- $v_k^{\max}$: local variance of the local smoothness estimator
- m_{k+1} is chosen to control the variability introduced by the previous stepsize $\eta_{k+1} \propto \bar{L}_k^{-1}$

Adaptive minibatch m_{k+1} for iterate update

Minibatch in parameter free setting

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{v_k^{\max}, \frac{\sigma_k^2}{D_0^2}\right\}\right\rceil\right)$$

Local $\bar{L}_k, \sigma_k^2, v_k^{\max} := \max_{0 \le i \le k} v_i$

Minibatch AS-SA Ghadimi & Lan 2016

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{L^2} \cdot \frac{\sigma^2}{D_0^2}\right\rceil\right)$$

Global smoothness L, variance σ^2

- σ_k^2 : local variance of the stochastic gradient error
- $v_k^{\max}$: local variance of the local smoothness estimator
- m_{k+1} is chosen to control the variability introduced by the previous stepsize $\eta_{k+1} \propto \bar{L}_k^{-1}$
- $\bar{L}_k^{-1}$ is not unbiased! $\Longrightarrow$ $v_k^{\max}$ controls the bias of $\bar{L}_k^{-1}$.

Adaptive minibatch m_{k+1} for iterate update

Minibatch in parameter free setting

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{v_k^{\max}, \frac{\sigma_k^2}{D_0^2}\right\}\right\rceil\right)$$

Local $\bar{L}_k, \sigma_k^2, v_k^{\max} := \max_{0 \le i \le k} v_i$

Minibatch AS-SA Ghadimi & Lan 2016

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{L^2} \cdot \frac{\sigma^2}{D_0^2}\right\rceil\right)$$

Global smoothness L, variance σ^2

- σ_k^2 : local variance of the stochastic gradient error
- $v_k^{\max}$: local variance of the local smoothness estimator
- m_{k+1} is chosen to control the variability introduced by the previous stepsize $\eta_{k+1} \propto \bar{L}_k^{-1}$
- $\bar{L}_k^{-1}$ is not unbiased! $\Longrightarrow$ $v_k^{\max}$ controls the bias of $\bar{L}_k^{-1}$.
- m_{k+1} is determined by previously observed quantities: $v_k^{\max}$, $\bar{L}_k$, and σ_k^2 — Adaptivive!

Adaptive minibatch m_{k+1} for iterate update

Minibatch in parameter free setting

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{v_k^{\max}, \frac{\sigma_k^2}{D_0^2}\right\}\right\rceil\right)$$

Local $\bar{L}_k, \sigma_k^2, v_k^{\max} := \max_{0\le i\le k} v_i$

Minibatch AS-SA Ghadimi & Lan 2016

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{L^2} \cdot \frac{\sigma^2}{D_0^2}\right\rceil\right)$$

Global smoothness L, variance σ^2

- σ_k^2 : local variance of the stochastic gradient error
- $v_k^{\max}$: local variance of the local smoothness estimator
- m_{k+1} is chosen to control the variability introduced by the previous stepsize $\eta_{k+1} \propto \bar{L}_k^{-1}$
- $\bar{L}_k^{-1}$ is not unbiased! $\Longrightarrow$ $v_k^{\max}$ controls the bias of $\bar{L}_k^{-1}$.
- m_{k+1} is determined by previously observed quantities: $v_k^{\max}$, $\bar{L}_k$, and σ_k^2 — Adaptivive!

For the next iteration: compute n_{k+1}

Adaptive minibatch n_{k+1} for stepsize update

Stochastic AC-FGM

$$n_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{ v_k^{\max}, \frac{\tilde{\sigma}_k^2 + \tilde{\sigma}_{k+1}^2}{D_0^2} \right\} \right\rceil\right)$$

Local $\bar{L}_k, \tilde{\sigma}_{k+1}^2, v_k^{\max} := \max_{0 \le i \le k} v_i$

Adaptive minibatch n_{k+1} for stepsize update

Stochastic AC-FGM

$$n_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{ v_k^{\max}, \frac{\tilde{\sigma}_k^2 + \tilde{\sigma}_{k+1}^2}{D_0^2} \right\} \right\rceil\right)$$

Local $\bar{L}_k, \tilde{\sigma}_{k+1}^2, v_k^{\max} := \max_{0 \le i \le k} v_i$

> $\tilde{\sigma}_{k+1}^2$: after m_{k+1}, x_{k+1} can be computed, and thus $\tilde{\sigma}_{k+1}^2$ is known for n_{k+1}.

$$\mathbb{E}_\xi\left[\|G(x_k,\xi) - \nabla f(x_k)\|^2 \,\middle|\, \mathcal{G}_k\right] \le \tilde{\sigma}_k^2, \quad \tilde{\sigma}_k^2 \in \mathcal{G}_k := \sigma(x_1, \ldots, x_k).$$

Adaptive minibatch n_{k+1} for stepsize update

Stochastic AC-FGM

$$n_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{ v_k^{\max}, \frac{\tilde{\sigma}_k^2 + \tilde{\sigma}_{k+1}^2}{D_0^2} \right\} \right\rceil\right)$$

Local $\bar{L}_k, \tilde{\sigma}_{k+1}^2, v_k^{\max} := \max_{0 \le i \le k} v_i$

- $\tilde{\sigma}_{k+1}^2$: after m_{k+1}, x_{k+1} can be computed, and thus $\tilde{\sigma}_{k+1}^2$ is known for n_{k+1}.

$$\mathbb{E}_\xi\left[\|G(x_k,\xi) - \nabla f(x_k)\|^2 \,\middle|\, \mathcal{G}_k\right] \le \tilde{\sigma}_k^2, \quad \tilde{\sigma}_k^2 \in \mathcal{G}_k := \sigma(x_1, \ldots, x_k).$$

- n_{k+1} is determined by previously observed quantities: $v_k^{\max}$, $\bar{L}_k$, and σ_k^2 — Adaptivive!

Adaptive minibatch n_{k+1} for stepsize update

Stochastic AC-FGM

$$n_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{v_k^{\max}, \frac{\tilde{\sigma}_k^2 + \tilde{\sigma}_{k+1}^2}{D_0^2}\right\}\right\rceil\right)$$

Local $\bar{L}_k, \tilde{\sigma}_{k+1}^2, v_k^{\max} := \max_{0 \le i \le k} v_i$

- $\tilde{\sigma}_{k+1}^2$: after m_{k+1}, x_{k+1} can be computed, and thus $\tilde{\sigma}_{k+1}^2$ is known for n_{k+1}.

$$\mathbb{E}_\xi\left[\|G(x_k,\xi) - \nabla f(x_k)\|^2 \,\middle|\, \mathcal{G}_k\right] \le \tilde{\sigma}_k^2, \quad \tilde{\sigma}_k^2 \in \mathcal{G}_k := \sigma(x_1, \ldots, x_k).$$

- n_{k+1} is determined by previously observed quantities: $v_k^{\max}$, $\bar{L}_k$, and σ_k^2 — Adaptivive!
- n_{k+1} is used to control the variability induced by the previous stepsize $\eta_{k+1} \propto \bar{L}_k^{-1}$

Adaptive minibatch n_{k+1} for stepsize update

Stochastic AC-FGM

$$n_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{ v_k^{\max}, \frac{\tilde{\sigma}_k^2 + \tilde{\sigma}_{k+1}^2}{D_0^2} \right\} \right\rceil\right)$$

Local $\bar{L}_k, \tilde{\sigma}_{k+1}^2, v_k^{\max} := \max_{0 \le i \le k} v_i$

- $\tilde{\sigma}_{k+1}^2$: after m_{k+1}, x_{k+1} can be computed, and thus $\tilde{\sigma}_{k+1}^2$ is known for n_{k+1}.

$$\mathbb{E}_\xi\left[\|G(x_k,\xi) - \nabla f(x_k)\|^2 \,\middle|\, \mathcal{G}_k\right] \le \tilde{\sigma}_k^2, \quad \tilde{\sigma}_k^2 \in \mathcal{G}_k := \sigma(x_1, \ldots, x_k).$$

- n_{k+1} is determined by previously observed quantities: $v_k^{\max}$, $\bar{L}_k$, and σ_k^2 — Adaptivive!
- n_{k+1} is used to control the variability induced by the previous stepsize $\eta_{k+1} \propto \bar{L}_k^{-1}$
- $\bar{L}_k^{-1}$ is not unbiased! $\Longrightarrow$ $v_k^{\max}$ controls the bias of $\bar{L}_k^{-1}$.

Three Batch Types & Filtration

Gradient Est.

Samples $\{\xi_{k,i}\}_{i=1}^{m_k}$

Compute G_k z_k, x_k, y_k

Stepsize update. I

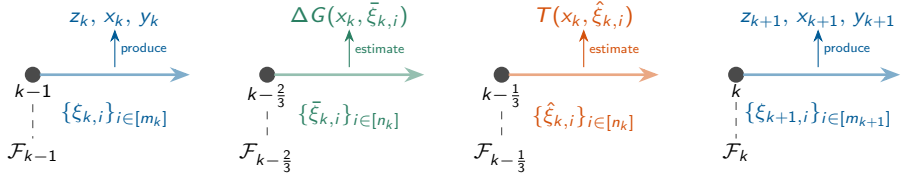
Samples $\{\bar{\xi}_{k,i}\}$

Compute ΔG

Stepsize update. II

Samples $\{\hat{\xi}_{k,i}\}$

Compute T



Three Batch Types & Filtration

Gradient Est.

Samples $\{\xi_{k,i}\}_{i=1}^{m_k}$

Compute G_k z_k, x_k, y_k

Stepsize update. I

Samples $\{\bar{\xi}_{k,i}\}$

Compute ΔG

Stepsize update. II

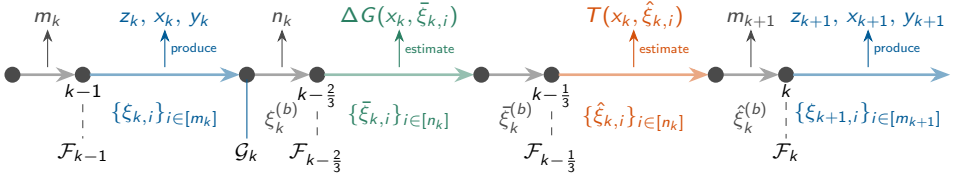
Samples $\{\hat{\xi}_{k,i}\}$

Compute T

Variance Est.

r_k samples

Update m_k, n_k



Type III: minibatch update batches

Minibatch in parameter free setting (unknown var)

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{\widehat{v}_k^{\max}, \frac{\widehat{\sigma}_k^2}{D_0^2}\right\}\right\rceil\right)$$

Local $\bar{L}_k, \widehat{\sigma}_k^2, \widehat{v}_k^{\max} := \max_{0\le i\le k} \widehat{v}_i$

Minibatch in parameter free setting (known var)

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{v_k^{\max}, \frac{\sigma_k^2}{D_0^2}\right\}\right\rceil\right)$$

Local $\bar{L}_k, \sigma_k^2, v_k^{\max} := \max_{0\le i\le k} v_i$

Type III: minibatch update batches

Minibatch in parameter free setting (unknown var)

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{\widehat{v}_k^{\max}, \frac{\widehat{\sigma}_k^2}{D_0^2}\right\}\right\rceil\right)$$

Local $\bar{L}_k, \widehat{\sigma}_k^2, \widehat{v}_k^{\max} := \max_{0\le i\le k} \widehat{v}_i$

Minibatch in parameter free setting (known var)

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{v_k^{\max}, \frac{\sigma_k^2}{D_0^2}\right\}\right\rceil\right)$$

Local $\bar{L}_k, \sigma_k^2, v_k^{\max} := \max_{0\le i\le k} v_i$

- The third type $\{\xi_{k,i}^b\}_{i=1}^{r_k} \Rightarrow \widehat{\sigma}_k$, $\{\bar{\xi}_{k,i}^b\}_{i=1}^{r_k} \Rightarrow \widehat{\delta}_k$, and $\{\widehat{\xi}_{k,i}^b\}_{i=1}^{r_k} \Rightarrow \widehat{v}_k$ fresh batches estimate variance without sacrificing adaptivity or acceleration.

Type III: minibatch update batches

Minibatch in parameter free setting (unknown var)

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{\widehat{v}_k^{\max}, \frac{\widehat{\sigma}_k^2}{D_0^2}\right\}\right\rceil\right)$$

Local $\bar{L}_k, \widehat{\sigma}_k^2, \widehat{v}_k^{\max} := \max_{0\le i\le k} \widehat{v}_i$

Minibatch in parameter free setting (known var)

$$m_{k+1} = \mathcal{O}\left(\left\lceil \frac{N(k+1)^2}{\bar{L}_k^2} \max\left\{v_k^{\max}, \frac{\sigma_k^2}{D_0^2}\right\}\right\rceil\right)$$

Local $\bar{L}_k, \sigma_k^2, v_k^{\max} := \max_{0\le i\le k} v_i$

- The third type $\{\xi_{k,i}^b\}_{i=1}^{r_k} \Rightarrow \widehat{\sigma}_k$, $\{\bar{\xi}_{k,i}^b\}_{i=1}^{r_k} \Rightarrow \widehat{\delta}_k$, and $\{\widehat{\xi}_{k,i}^b\}_{i=1}^{r_k} \Rightarrow \widehat{v}_k$ fresh batches estimate variance without sacrificing adaptivity or acceleration.
- The framework is estimator-agnostic: any good filtration-adapted variance estimator can be used.

Assumptions I: unbiased estimate

A1.Conditional unbiasedness. Given x_k for the main update samples ξ_{k+1},

$$\mathbb{E}_{\xi_{k+1}}[G(x_k, \xi_{k+1}) \mid \mathcal{F}_k] = \nabla f(x_k). \quad \text{classic}$$

Assumptions I: unbiased estimate

A1.Conditional unbiasedness. Given x_k for the main update samples ξ_{k+1},

$$\mathbb{E}_{\xi_{k+1}}[G(x_k, \xi_{k+1}) \mid \mathcal{F}_k] = \nabla f(x_k). \quad \text{classic}$$

For the stepsize selection samples $\bar{\xi}_k$ and $\widehat{\xi}_k$, it holds that

$$\mathbb{E}_{\bar{\xi}_k}\left[G(x_k, \bar{\xi}_k) \mid \mathcal{F}_{k-\frac{2}{3}}\right] = \nabla f(x_k), \quad \mathbb{E}_{\widehat{\xi}_k}\left[G(x_k, \widehat{\xi}_k) \mid \mathcal{F}_{k-\frac{1}{3}}\right] = \nabla f(x_k),$$
$$\mathbb{E}_{\widehat{\xi}_k}\left[F(x_k, \widehat{\xi}_k) \mid \mathcal{F}_{k-\frac{1}{3}}\right] = f(x_k), \quad \mathbb{E}_{\widehat{\xi}_k}\left[F(x_{k-1}, \widehat{\xi}_k) \mid \mathcal{F}_{k-\frac{1}{3}}\right] = f(x_{k-1}).$$

Assumptions I: unbiased estimate

A1.Conditional unbiasedness. Given x_k for the main update samples ξ_{k+1},

$$\mathbb{E}_{\xi_{k+1}}[G(x_k, \xi_{k+1}) \mid \mathcal{F}_k] = \nabla f(x_k). \quad \text{classic}$$

For the stepsize selection samples $\bar{\xi}_k$ and $\widehat{\xi}_k$, it holds that

$$\mathbb{E}_{\bar{\xi}_k}\left[G(x_k, \bar{\xi}_k) \mid \mathcal{F}_{k-\frac{2}{3}}\right] = \nabla f(x_k), \quad \mathbb{E}_{\widehat{\xi}_k}\left[G(x_k, \widehat{\xi}_k) \mid \mathcal{F}_{k-\frac{1}{3}}\right] = \nabla f(x_k),$$
$$\mathbb{E}_{\widehat{\xi}_k}\left[F(x_k, \widehat{\xi}_k) \mid \mathcal{F}_{k-\frac{1}{3}}\right] = f(x_k), \quad \mathbb{E}_{\widehat{\xi}_k}\left[F(x_{k-1}, \widehat{\xi}_k) \mid \mathcal{F}_{k-\frac{1}{3}}\right] = f(x_{k-1}).$$

Classical unbiasedness extends naturally to the new filtrations induced by the parameter-free setting.

Assumptions II: Bounded local conditional variance

A2. Bounded local conditional variance. Given x_k, $\exists\ \sigma_k \geq 0$ s.t. for a fresh sample ξ_{k+1},

$$\mathbb{E}_{\xi_{k+1}}\left[\|G(x_k, \xi_{k+1}) - \nabla f(x_k)\|^2 \,\middle|\, \mathcal{F}_k\right] \leq \sigma_k^2. \quad \text{classic}$$

Assumptions II: Bounded local conditional variance

A2. Bounded local conditional variance. Given x_k, $\exists\ \sigma_k \geq 0$ s.t. for a fresh sample ξ_{k+1},

$$\mathbb{E}_{\xi_{k+1}}\left[\|G(x_k, \xi_{k+1}) - \nabla f(x_k)\|^2 \,\middle|\, \mathcal{F}_k\right] \leq \sigma_k^2. \quad \text{classic}$$

$\exists\ \delta_k \geq 0$ such that, for fresh batch-size and stepsize selection samples ξ_k^b and $\bar{\xi}_k$,

$$\mathbb{E}_{\xi_k^b}\left[\|G(x_k, \xi_k^b) - \nabla f(x_k)\|^2 \,\middle|\, \mathcal{G}_k\right] \leq \delta_k^2,$$
$$\mathbb{E}_{\bar{\xi}_k}\left[\|G(x_k, \bar{\xi}_k) - \nabla f(x_k)\|^2 \,\middle|\, \mathcal{G}_k\right] \leq \delta_k^2.$$

Assumptions II: Bounded local conditional variance

A2. Bounded local conditional variance. Given x_k, $\exists\ \sigma_k \geq 0$ s.t. for a fresh sample ξ_{k+1},

$$\mathbb{E}_{\xi_{k+1}}\left[\|G(x_k,\xi_{k+1}) - \nabla f(x_k)\|^2 \,\middle|\, \mathcal{F}_k\right] \leq \sigma_k^2. \quad \text{classic}$$

$\exists\, \delta_k \geq 0$ such that, for fresh batch-size and stepsize selection samples ξ_k^b and $\bar{\xi}_k$,

$$\mathbb{E}_{\xi_k^b}\left[\|G(x_k,\xi_k^b) - \nabla f(x_k)\|^2 \,\middle|\, \mathcal{G}_k\right] \leq \delta_k^2,$$
$$\mathbb{E}_{\bar{\xi}_k}\left[\|G(x_k,\bar{\xi}_k) - \nabla f(x_k)\|^2 \,\middle|\, \mathcal{G}_k\right] \leq \delta_k^2.$$

$\exists v_k \geq 0$ such that for a fresh stepsize selection sample $\widehat{\xi}_k$,

$$\mathbb{E}_{\widehat{\xi}_k}\left[|\widetilde{L}_k(\widehat{\xi}_k) - L_k|^2 \,\middle|\, \mathcal{F}_{k-\frac{2}{3}}\right] \leq v_k^2,$$

where L_k is defined is the local smoothness estimator.

Assumptions II: Bounded local conditional variance

A2. Bounded local conditional variance. Given x_k, $\exists\ \sigma_k \geq 0$ s.t. for a fresh sample ξ_{k+1},

$$\mathbb{E}_{\xi_{k+1}}\left[\|G(x_k,\xi_{k+1}) - \nabla f(x_k)\|^2 \,\middle|\, \mathcal{F}_k\right] \leq \sigma_k^2. \quad \text{classic}$$

$\exists\,\delta_k \geq 0$ such that, for fresh batch-size and stepsize selection samples ξ_k^b and $\bar{\xi}_k$,

$$\mathbb{E}_{\xi_k^b}\left[\|G(x_k,\xi_k^b) - \nabla f(x_k)\|^2 \,\middle|\, \mathcal{G}_k\right] \leq \delta_k^2,$$
$$\mathbb{E}_{\bar{\xi}_k}\left[\|G(x_k,\bar{\xi}_k) - \nabla f(x_k)\|^2 \,\middle|\, \mathcal{G}_k\right] \leq \delta_k^2.$$

$\exists v_k \geq 0$ such that for a fresh stepsize selection sample $\widehat{\xi}_k$,

$$\mathbb{E}_{\widehat{\xi}_k}\left[|\widetilde{L}_k(\widehat{\xi}_k) - L_k|^2 \,\middle|\, \mathcal{F}_{k-\frac{2}{3}}\right] \leq v_k^2,$$

where L_k is defined is the local smoothness estimator.

> Classical bounded variance extends naturally to the new filtrations induced by the parameter-free setting; v_k is due to the bias of stepsize estimator.

Assumptions III: Finite-sample cocoercivity–smoothness

A3. Given a query pair (x_{k-1}, x_k), $\exists\ \mathcal{L}(\bar{\xi}_k, \hat{\xi}_k),\ \mathcal{L}(\hat{\xi}_k) > 0$ s.t.

$$\frac{\|\Delta G(x_k, \bar{\xi}_k)\|^2}{2\mathcal{L}(\bar{\xi}_k, \hat{\xi}_k)} \ \leq\ T(x_k, \hat{\xi}_k) \ \leq\ \frac{\mathcal{L}(\hat{\xi}_k)}{2}\|x_k - x_{k-1}\|^2, \quad \text{a.s.}$$

Assumptions III: Finite-sample cocoercivity–smoothness

A3. Given a query pair (x_{k-1}, x_k), $\exists\ \mathcal{L}(\bar{\xi}_k, \hat{\xi}_k),\ \mathcal{L}(\hat{\xi}_k) > 0$ s.t.

$$\frac{\|\Delta G(x_k, \bar{\xi}_k)\|^2}{2\mathcal{L}(\bar{\xi}_k, \hat{\xi}_k)} \ \leq\ T(x_k, \hat{\xi}_k) \ \leq\ \frac{\mathcal{L}(\hat{\xi}_k)}{2}\|x_k - x_{k-1}\|^2, \quad \text{a.s.}$$

A finite-sample analogue of cocoercivity + smoothness. Does not require each $F(\cdot, \hat{\xi})$ to be convex or smooth — only locally well-behaved along (x_{k-1}, x_k).

Adaptivity to the Lipschitz constant

Theorem 1 (Ji & Lan, 2026)

Under A1–A3, with the previously defined adaptive stepsize and minibatches rules, to reach ε solution such that $\mathbb{E}[\Phi(x_N) - \Phi(x^*)] \leq \varepsilon$, stochastic AC-FGM requires

$$\mathcal{O}\left(\sqrt{\frac{\mathcal{L}D_0^2}{\varepsilon} \cdot \max\left\{\frac{v_{\max}}{v_0}, 1\right\}}\right) \quad \text{iterations.}$$

Adaptivity to the Lipschitz constant

Theorem 1 (Ji & Lan, 2026)

Under A1–A3, with the previously defined adaptive stepsize and minibatches rules, to reach ε solution such that $\mathbb{E}[\Phi(x_N) - \Phi(x^*)] \leq \varepsilon$, stochastic AC-FGM requires

$$\mathcal{O}\left(\sqrt{\frac{\mathcal{L}D_0^2}{\varepsilon} \cdot \max\left\{\frac{v_{\max}}{v_0}, 1\right\}}\right) \quad \text{iterations.}$$

- Matches the L-known AC-SA rate Lan (2012) — optimal in ε Nemirovski (1985)

Adaptivity to the Lipschitz constant

Theorem 1 (Ji & Lan, 2026)

Under A1–A3, with the previously defined adaptive stepsize and minibatches rules, to reach ε solution such that $\mathbb{E}[\Phi(x_N) - \Phi(x^*)] \leq \varepsilon$, stochastic AC-FGM requires

$$\mathcal{O}\left(\sqrt{\frac{\mathcal{L}D_0^2}{\varepsilon} \cdot \max\left\{\frac{v_{\max}}{v_0}, 1\right\}}\right) \quad \text{iterations.}$$

- Matches the L-known AC-SA rate Lan (2012) — optimal in ε Nemirovski (1985)
- Pessimistic $\mathcal{L}$, $v_{\max}$, due to the randomness of the stepsize when seeking in expectation guarantee; see also Lan, Li, & Xu 2024;

Adaptivity to the Lipschitz constant

Theorem 1 (Ji & Lan, 2026)

Under A1–A3, with the previously defined adaptive stepsize and minibatches rules, to reach ε solution such that $\mathbb{E}[\Phi(x_N) - \Phi(x^*)] \leq \varepsilon$, stochastic AC-FGM requires

$$\mathcal{O}\left(\sqrt{\frac{\mathcal{L}D_0^2}{\varepsilon} \cdot \max\left\{\frac{v_{\max}}{v_0}, 1\right\}}\right) \quad \text{iterations.}$$

- Matches the L-known AC-SA rate Lan (2012) — optimal in ε Nemirovski (1985)
- Pessimistic $\mathcal{L}, v_{\max}$, due to the randomness of the stepsize when seeking in expectation guarantee; see also Lan, Li, & Xu 2024;
- Under light-tail assumptions, $\mathcal{L}, v_{\max} \to$ local quantities $\widehat{L}_N$ and $v_N^{\max}$.

Sample complexity

To reach ε solution s.t. $\mathbb{E}[\Phi(x_N) - \Phi(x^*)] \le \varepsilon$, stochastic AC-FGM requires

$$\mathcal{O}\left(\sqrt{\frac{\mathcal{L}D_0^2}{\varepsilon} \cdot \frac{v_{\max}}{v_0}} + \frac{r_N \mathcal{L}^2 D_0^2}{\varepsilon^2} \cdot \frac{v_{\max}^2}{v_0^2}\right) \quad \text{samples}$$

where r_N is the *average variance-to-smoothness ratio* along the trajectory

$$r_N^2 := \frac{1}{N}\sum_{k=1}^{N} \frac{v_{k-1}^{\max} D_0^2 + \sigma_{k-1}^2}{L_k^2} \Longrightarrow \frac{\sigma^2}{L^2}$$

Sample complexity

To reach ε solution s.t. $\mathbb{E}[\Phi(x_N) - \Phi(x^*)] \leq \varepsilon$, stochastic AC-FGM requires

$$\mathcal{O}\left(\sqrt{\frac{\mathcal{L}D_0^2}{\varepsilon} \cdot \frac{v_{\max}}{v_0}} + \frac{r_N \mathcal{L}^2 D_0^2}{\varepsilon^2} \cdot \frac{v_{\max}^2}{v_0^2}\right) \quad \text{samples}$$

where r_N is the *average variance-to-smoothness ratio* along the trajectory

$$r_N^2 := \frac{1}{N}\sum_{k=1}^{N} \frac{v_{k-1}^{\max} D_0^2 + \sigma_{k-1}^2}{L_k^2} \Longrightarrow \frac{\sigma^2}{L^2}$$

To reach ε solution s.t. $\mathbb{E}[\Phi(x_N) - \Phi(x^*)] \leq \varepsilon$, AC-SA Lan 2012 requires

$$\mathcal{O}\left(\sqrt{\frac{LD_0^2}{\varepsilon}} + \frac{\sigma^2 D_0^2}{\varepsilon^2}\right) \quad \text{samples}$$

Iteration limit N free: anchored regularization

The proximal subproblem includes an extra anchor term:

$$z_k = \operatorname*{argmin}_{z \in X} \left\{ \langle G_k, z \rangle + h(z) + \tfrac{1}{2\eta_k} \|y_{k-1} - z\|^2 + + \frac{\gamma_k}{2\eta_k} \|y_0 - z\|^2 \right\}$$

Iteration limit N free: anchored regularization

The proximal subproblem includes an extra anchor term:

$$z_k = \operatorname*{argmin}_{z \in X} \left\{ \langle G_k, z \rangle + h(z) + \tfrac{1}{2\eta_k} \|y_{k-1} - z\|^2 + + \frac{\gamma_k}{2\eta_k} \|y_0 - z\|^2 \right\}$$

- γ_k chosen small (decaying) so convergence rate is unchanged

Iteration limit N free: anchored regularization

The proximal subproblem includes an extra anchor term:

$$z_k = \operatorname*{argmin}_{z \in X} \left\{ \langle G_k, z \rangle + h(z) + \tfrac{1}{2\eta_k} \|y_{k-1} - z\|^2 + + \frac{\gamma_k}{2\eta_k} \|y_0 - z\|^2 \right\}$$

- γ_k chosen small (decaying) so convergence rate is unchanged
- Enables horizon-free analysis — no prior knowledge of total iterations N

Adaptive to the iteration limit

Theorem 2 (Ji & Lan, 2026)

Under A1–A3, with anchored regularization $\gamma_k = \frac{1}{k}$, $\tau_k = \frac{k+2-\beta}{2}$, and suppose $\eta_1 > 0$ and

$$\eta_2 = \min\left\{\frac{1}{16\bar{L}_1}, \frac{2(1-\beta)}{3-\beta}\eta_1\right\}, \quad \eta_k = \min\left\{\frac{k-1}{16\bar{L}_{k-1}}, \frac{(k-1)(k+2-\beta)}{k^2}\eta_{k-1}\right\}, \quad \forall\, k \geq 3.$$

Furthermore, suppose the mini batch size satisfies

$$m_k = \mathcal{O}\left(\left\lceil \frac{(k+1)\,k^2}{\bar{L}_{k-1}^2}\max\left\{v_{k-1}^{\max}, \frac{\sigma_{k-1}^2}{D_0^2}\right\}\right\rceil\right), \; n_k = \mathcal{O}\left(\left\lceil \frac{(k+1)\,k^2}{\bar{L}_{k-1}^2}\max\left\{v_{k-1}^{\max}, \frac{\delta_k^2}{D_0^2}\right\}\right\rceil\right).$$

To reach ε solution such that $\mathbb{E}[\Phi(x_N) - \Phi(x^*)] \leq \varepsilon$ on A, stochastic AC-FGM requires

$$\mathcal{O}\left(\sqrt{\frac{\mathcal{L}D_0^2}{\varepsilon}\cdot\max\left\{\frac{v_{\max}}{v_0}, 1\right\}}\right) \quad \text{iterations.}$$

Adaptivity to the local variances

Theorem 3 (Ji & Lan, 2026)

Under A1-A3, the same parameter choices as Theorem 2, suppose

$$A := \{\forall\, k, \widehat{v}_{k-1} \geq v_{k-1}, \widehat{\sigma}^2_{k-1} \geq \sigma^2_{k-1}, \widehat{\delta}^2_k \geq \delta^2_k\}, \quad \mathbb{P}(A) \geq 1 - \delta.$$

Furthermore, suppose the mini batch size satisfies

$$m_k = \mathcal{O}\left(\left\lceil \frac{(k+1)\,k^2}{\bar{L}^2_{k-1}} \max\left\{\widehat{v}^{\max}_{k-1}, \frac{\widehat{\sigma}^2_{k-1}}{D_0^2}\right\}\right\rceil\right),\ n_k = \mathcal{O}\left(\left\lceil \frac{(k+1)\,k^2}{\bar{L}^2_{k-1}} \max\left\{\widehat{v}^{\max}_{k-1}, \frac{\widehat{\delta}^2_k}{D_0^2}\right\}\right\rceil\right).$$

To reach ε solution such that $\mathbb{E}[\Phi(x_N) - \Phi(x^*)] \leq \varepsilon$ on A, stochastic AC-FGM requires

$$\mathcal{O}\left(\sqrt{\frac{\mathcal{L} D_0^2}{\varepsilon} \cdot \max\left\{\frac{v_{\max}}{v_0}, 1\right\}}\right) \quad \text{iterations.}$$

High-Probability Guarantee

Theorem 4 (Ji & Lan, 2026)

Under additional light tail assumption, w.p. at least $1-(N+1)e^{-\Lambda^2/3}-4(N+1)e^{-\Lambda}$:

$$\Phi(x_N)-\Phi(x^*) \le \varepsilon \quad \text{in}$$

$$\mathcal{O}\left(\sqrt{\frac{\widehat{L}_N D_0^2}{\varepsilon}\cdot\max\left\{\frac{v_{N+1}^{\max}}{v_0},1\right\}}\right) \quad \text{iterations.}$$

High-Probability Guarantee

Theorem 4 (Ji & Lan, 2026)

Under additional light tail assumption, w.p. at least $1-(N+1)e^{-\Lambda^2/3}-4(N+1)e^{-\Lambda}$:

$$\Phi(x_N)-\Phi(x^*) \le \varepsilon \quad \text{in}$$

$$\mathcal{O}\left(\sqrt{\frac{\widehat{L}_N D_0^2}{\varepsilon}\cdot\max\left\{\frac{v_{N+1}^{\max}}{v_0},1\right\}}\right) \quad \text{iterations.}$$

- Trajectory-dependent: $\widehat{L}_N$, $v_{N+1}^{\max}$ depend only on iterates *actually visited*
- Prior work (Kreisler et al. 2024): bound depends on $\sigma^* = \max_{x\in\mathbb{B}(x^*,2D_0)} \sigma_x$ — supremum over entire ball
- Ours: no global worst-case penalty; pays only for noise and curvature encountered

High-Probability sample complexity

Stochastic AC-FGM requires

$$\mathcal{O}\left(\sqrt{\frac{L_N D_0^2}{\varepsilon}\cdot\frac{v_{N+1}^{\max}}{v_0}}+\frac{r_N\mathcal{L}^2 D_0^2}{\varepsilon^2}\cdot\frac{(v_{N+1}^{\max})^2}{v_0^2}\right) \quad \text{samples}$$

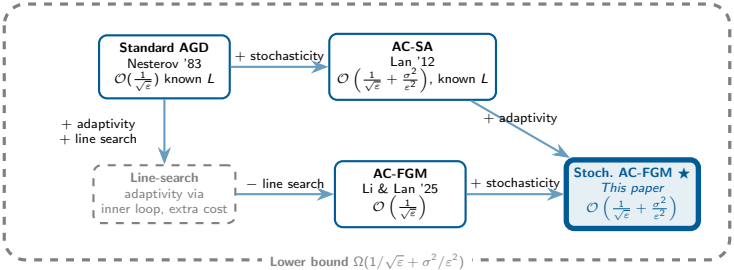
where r_N is the *average variance-to-smoothness ratio* along the trajectory

$$r_N^2 := \frac{1}{N}\sum_{k=1}^{N}\frac{v_{k-1}^{\max}D_0^2+\sigma_{k-1}^2}{\bar{L}_k^2} \Longrightarrow \frac{\sigma^2}{L^2}$$

To reach ε solution s.t. $\mathbb{E}[\Phi(x_N)-\Phi(x^*)]\le\varepsilon$, AC-SA Lan 2012 requires

$$\mathcal{O}\left(\sqrt{\frac{LD_0^2}{\varepsilon}}+\frac{\sigma^2 D_0^2}{\varepsilon^2}\right) \quad \text{samples}$$

Take-Home Message



Stochastic AC-FGM achieves:

- Adaptive to the Lipschitz constant L, the iteration horizon N, and the local variances
- Optimal convergence rate without boundedness assumption

Thank You

Stochastic Auto-conditioned Fast Gradient Methods with Op

Yao Ji & Guanghui (George) Lan

Preprint: Stay tuned for arXiv soon!

Contact: yaoji@gatech.edu george.lan@isye.gatech.edu