A Short and Unified Convergence Analysis of the SAG, SAGA, and IAG Algorithms

Presenter: Feng Zhu fzhu5@ncsu.edu
Department of Electrical and Computer Engineering
North Carolina State University
May 31st, 2026
Advisors: Aritra Mitra and Robert W. Heath Jr.

NC STATE UNIVERSITY
Electrical and Computer Engineering

UC San Diego

Problem Setup

Goal: Solve the optimization problem $\min_{x \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^{N} f_i(x)$

- Each component function f_i is L-smooth
- Global function f is μ-strongly convex
- x^* is the minimizer of f

Iterative Approaches

- General update rule:

$$x_{k+1} = x_k - \alpha g_k, \qquad k = 0, \cdots, K-1$$

- *Gradient Descent (GD)*: $g_k = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x_k)$, **linear** convergence to x^*, but **high per-iteration computation cost** (N gradient evaluations)

- *Stochastic Gradient Descent (SGD)*: $g_k = f_{i_k}(x_k)$, light computation cost, but **high variance** and **sub-linear convergence**

Q: Can we bring down the **variance** of SGD?

Variance-Reduced Algorithms

Core Idea: Maintain memory of previously computed gradients, and update the parameter exploiting memory of past gradients of all components

- SAG (biased gradient)

$$g_k^{SAG} = \frac{1}{N}\sum_{i\in[N]} \nabla f_i(x_{\tau_{i,k}}) + \frac{\nabla f_{i_k}(x_k) - \nabla f_{i_k}\left(x_{\tau_{i_k,k}}\right)}{N}$$

- SAGA (unbiased gradient)

$$g_k^{SAGA} = \frac{1}{N}\sum_{i\in[N]} \nabla f_i(x_{\tau_{i,k}}) + \nabla f_{i_k}(x_k) - \nabla f_{i_k}\left(x_{\tau_{i_k,k}}\right)$$

- IAG (deterministic counterpart of SAG)
  - Sampling index i_k in a **deterministic** manner from $[N]$
- All achieve **linear** convergence rates

Proofs of SAG, SAGA, and IAG

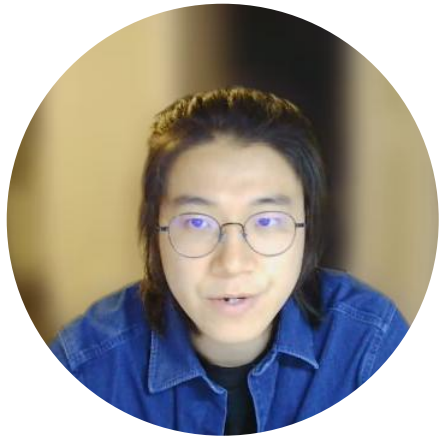
- Proof of SAG [1] is **notoriously challenging** and requires computer-aided analysis
- Proof of SAGA [2] is easier but **fails to explain** the convergence behavior of SAG
- Analysis of IAG [3] is **fundamentally different** from SAG/SAGA, and has a **much slower rate** than SAG/SAGA and GD.

Q: Can we **UNIFY** the proofs of SAG, SAGA and IAG?

[1] Schmidt, M., Le Roux, N., & Bach, F. (2017). *Minimizing finite sums with the stochastic average gradient.* Mathematical F
[2] Defazio, A., Bach, F., & Lacoste-Julien, S. (2014). *SAGA: A fast incremental gradient method with support for non-strongly c objectives.* NeurIPS
[3] Gurbuzbalaban, M., Ozdaglar, A., & Parrilo, P.A. (2017). *On the convergence rate of incremental aggregated gradient algorithm*

Contributions

- First unified proof of SAG, SAGA and IAG
- High-probability bounds for SAG and SAGA under both IID & Markovian sampling
- Significantly improve upon best-known bounds for IAG



Module 1 – Bounded Staleness

- Establish a high-probability event $\mathcal{G}$ where the gradient staleness of SAG/SAGA is bounded:

> **Lemma** (informal): *With high probability, the staleness of all component gradients are upper bounded by some constant* $\tau = \tilde{\mathcal{O}}(N)$.

- **Intuition:** Denote this event as $\mathcal{G}$. Conditioning on $\mathcal{G}$, we can treat SAG/SAGA as methods with ***uniformly bounded delay***.

Module 2 – Lyapunov function

- Design a simple *Lyapunov function*

$$V_k := f(x_k) - f(x^*) + L\alpha^2 W_k,$$

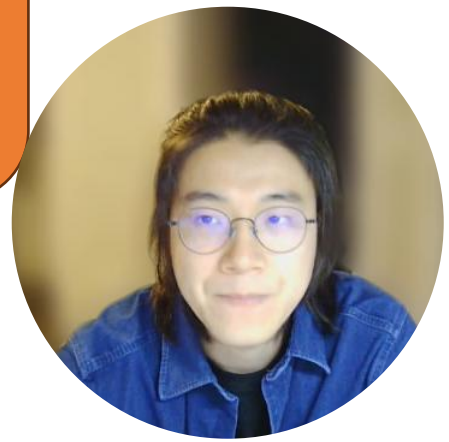
$$W_k := \sum_{j=1}^{\tau} (\tau - j + 1)\left\|g_{k-j}\right\|^2$$

- **Intuition:** Use a *shifted window* term to track the recent history

Module 1
High-Probability Bound on Staleness

Module 2
Novel Lyapunov Function to Control Staleness (Delays)

The two modules are completely DECOUPLED!



Main Result:

Theorem 1. *Given* $\delta \in (0,1), \tau = \left\lceil \frac{8N}{3} \log \frac{NK}{\delta} \right\rceil$ *and* $\alpha = \frac{1}{16L\tau}$, *the following holds for both SAG, SAGA with probability at least* $1 - \delta$ *for* $K > \tau$:

$$f(x_K) - f(x^*) \leq \mathcal{O}\left(\left(1 - \frac{1}{64\tau\kappa}\right)^K\right),$$

where $\kappa = L/\mu$.

Main Takeaways:

- Unified **linear** convergence rate for SAG/SAGA
- **First high-probability rate** for SAG/SAGA
- Result holds **deterministically** for IAG, with improvement from ***quadratic*** on κ **and** τ to ***linear*** dependence on both

Conclusion

- Introduced a *unified* **proof framework** yielding linear convergence rates for SAG, SAGA and IAG
- First high-probability bounds for SAG and SAGA
- Significantly sharpened rates for IAG
- Analysis framework extendible to *single-loop* VR algorithms with stochastic sub-sampling: bounded staleness -> specifically designed Lyapunov function

Thank you!