# Minimizing Finite Sums with the Stochastic Average Gradient

Mark Schmidt, Nicolas Le Roux, Francis Bach

Piyushi Manupriya
CS5660 Theory Paper Presentation

Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad; India

July 31, 2020

# Overview

# Introduction

- $N$: no of observations.

# Introduction

- $N$: no of observations.
- $P$: dimension of each observation.

# Introduction

- $N$: no of observations.
- $P$: dimension of each observation.
- **Objective:** To minimize the sum of finite set of **Smooth functions**.
  $\min_{x \in \mathbb{R}^P} g(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(x)$

# Introduction

- $N$: no of observations.
- $P$: dimension of each observation.
- **Objective:** To minimize the sum of finite set of **Smooth functions**.
  $\min_{x \in \mathbb{R}^P} g(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(x)$
  - To efficiently solve when $N$ or $P$ is very large.

# Introduction

- $N$: no of observations.
- $P$: dimension of each observation.
- **Objective:** To minimize the sum of finite set of **Smooth functions**. $\min_{x \in \mathbb{R}^P} g(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(x)$
  - To efficiently solve when $N$ or $P$ is very large.
- Assumptions:
  - g is $\mu-$Strongly convex. $g(y) \geq g(x) + g'(x)^T(y-x) + \frac{\mu}{2}||y-x||^2$
  - $f_i$ Convex. $f_i(y) - f_i(x) \geq f_i'(x)^T(y-x)$
  - $f_i'$ is $L$-Lipschitz continuous. $|f_i'(y) - f_i'(x)| \leq L||y-x||$

# Introduction

- Deterministic gradient method:
  - $x_{t+1} = x_t - \frac{\alpha_t}{N} \sum_{i=1}^{N} f_i'(x_t)$
  - **Linear Convergence** rate: $O(\rho^t)$
  - Iteration cost is linear in N

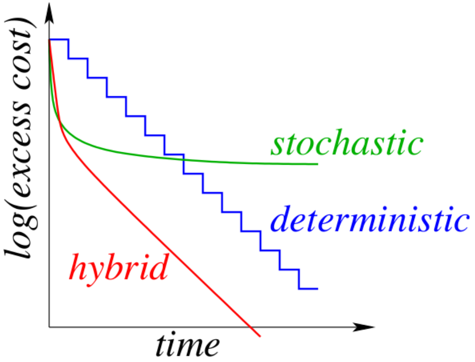
# Introduction

- Deterministic gradient method:
  - $x_{t+1} = x_t - \frac{\alpha_t}{N} \sum_{i=1}^{N} f_i'(x_t)$
  - **Linear Convergence** rate: $O(\rho^t)$
  - Iteration cost is linear in N
- Stochastic gradient method:
  - Random selection of $i(t)$ from $\{1, 2, ..., N\}$ $x_{t+1} = x_t - \alpha_t f_{i(t)}'(x_t)$
  - Iteration cost is indpendent of N.
  - **Sublinear convergence**: $O(\frac{1}{t})$
  - Needs carefully chosen step-size.

# Introduction

- Deterministic gradient method:
  - $x_{t+1} = x_t - \frac{\alpha_t}{N} \sum_{i=1}^{N} f_i'(x_t)$
  - **Linear Convergence** rate: $O(\rho^t)$
  - Iteration cost is linear in N
- Stochastic gradient method:
  - Random selection of $i(t)$ from $\{1, 2, ..., N\}$ $x_{t+1} = x_t - \alpha_t f_{i(t)}'(x_t)$
  - Iteration cost is indpendent of N.
  - **Sublinear convergence**: $O(\frac{1}{t})$
  - Needs carefully chosen step-size.
- **Paper Objective:** $O(\rho^t)$ rate with $O(1)$ iteration cost with constant step-size.

- FG method has $O(N)$ cost with $O(\rho^t)$ rate.
- SG method has $O(1)$ cost with $O(1/t)$ rate.



# Proposed SAG Algorithm

- SAG Iterations:
  - Sample a random index $i_k$ from $\{1, \ldots N\}$

# Proposed SAG Algorithm

- SAG Iterations:
  - Sample a random index $i_k$ from $\{1, ...N\}$
  - $x^{k+1} = x^k - \frac{\alpha_k}{N} \sum_{i=1}^{N} y_i^k$

# Proposed SAG Algorithm

- SAG Iterations:
  - Sample a random index $i_k$ from $\{1, ...N\}$
  - $x^{k+1} = x^k - \frac{\alpha_k}{N} \sum_{i=1}^{N} y_i^k$
  - $y_i^k = f_i'(x^k)$ if $i = i_k$

# Proposed SAG Algorithm

- SAG Iterations:
  - Sample a random index $i_k$ from $\{1, ...N\}$
  - $x^{k+1} = x^k - \frac{\alpha_k}{N} \sum_{i=1}^{N} y_i^k$
  - $y_i^k = f_i'(x^k)$ if $i = i_k$
    $y_i^k = y_i^{k-1}$ otherwise. (gradient at the last iteration where $i$ was selected)

# Proposed SAG Algorithm

- SAG Iterations:
  - Sample a random index $i_k$ from $\{1, ...N\}$
  - $x^{k+1} = x^k - \frac{\alpha_k}{N} \sum_{i=1}^{N} y_i^k$
  - $y_i^k = f_i'(x^k)$ if $i = i_k$
    $y_i^k = y_i^{k-1}$ otherwise. (gradient at the last iteration where $i$ was selected)
- Needs storage of $N$ gradient vectors. $O(NP)$ storage.

# Notations

$$e = \begin{pmatrix} I \\ \vdots \\ I \end{pmatrix} \in \mathbb{R}^{np \times p}, \quad f'(x) = \begin{pmatrix} f_1'(x) \\ \vdots \\ f_n'(x) \end{pmatrix} \in \mathbb{R}^{np}, \quad \theta^k = \begin{pmatrix} y_1^k \\ \vdots \\ y_n^k \\ x^k \end{pmatrix} \in \mathbb{R}^{(n+1)p}, \quad \theta^* = \begin{pmatrix} f_1'(x^*) \\ \vdots \\ f_n'(x^*) \\ x^* \end{pmatrix}$$

$$\theta^* \in \mathbb{R}^{(n+1)p}$$

$$e = \begin{pmatrix} I \\ \vdots \\ I \end{pmatrix} \in \mathbb{R}^{np \times p}, \quad f'(x) = \begin{pmatrix} f_1'(x) \\ \vdots \\ f_n'(x) \end{pmatrix} \in \mathbb{R}^{np}, \quad \theta^k = \begin{pmatrix} y_1^k \\ \vdots \\ y_n^k \\ x^k \end{pmatrix} \in \mathbb{R}^{(n+1)p}, \quad \theta^* = \begin{pmatrix} f_1'(x^*) \\ \vdots \\ f_n'(x^*) \\ x^* \end{pmatrix}$$

$\theta^* \in \mathbb{R}^{(n+1)p}$

- If $P = \begin{pmatrix} A & b \\ b^T & c \end{pmatrix}$, for $A \in \mathbb{R}^{np \times np}$, $b \in \mathbb{R}^{np \times p}$, $c \in \mathbb{R}^{p \times p}$,

then $(\theta^k - \theta^*)^T \begin{pmatrix} A & b \\ b^T & c \end{pmatrix} (\theta^k - \theta^*)$

$$= (y^k - f'(x^*))^T A (y^k - f'(x^*)) + 2(y^k - f'(x^*))^T b (x^k - x^*) + (x^k - x^*)^T c (x^k - x^*)$$

# Notations

$y_i^k = f_i'(x^k)$ if $i = i_k$

# Notations

$y_i^k = f_i'(x^k)$ if $i = i_k$
$y_i^k = y_i^{k-1}$ otherwise.

# Notations

$y_i^k = f_i'(x^k)$ if $i = i_k$
$y_i^k = y_i^{k-1}$ otherwise.

- $y_i^k = \left(1-\frac{1}{n}\right) y_i^{k-1} + \frac{1}{n} f_i'(x^{k-1}) + z_i^k \left[ f_i'(x^{k-1}) - y_i^{k-1} \right]$ — ①

where $z_i^k$ is a R.V., takes value $\left(1-\frac{1}{n}\right)$ with probability $\frac{1}{n}$
& $\left(-\frac{1}{n}\right)$ with probability $\left(1-\frac{1}{n}\right)$

REASON: $P_r$ (sampled $i_k = i$) $= \frac{1}{n}$ & when $i_k = i$, we want $y_i^k = f_i'(x^{k-1})$

① says that with probability $\frac{1}{n}$; $z_i^k$ takes $\left(1-\frac{1}{n}\right)$ making

$y_i^k = f_i'(x^{k-1})$

# Notations

$$x^{k+1} = x^k - \frac{\alpha_k}{N} \sum_{i=1}^{N} y_i^k$$

# Notations

$x^{k+1} = x^k - \frac{\alpha_k}{N} \sum_{i=1}^{N} y_i^k$

- Similarly, $x^k = x^{k-1} - \frac{\alpha}{n} \sum_{i=1}^{n} \left[ \left(1 - \frac{1}{n}\right) y_i^{k-1} + \frac{1}{n} f_i'(x^{k-1}) + z_i^k \left[ f_i'(x^{k-1}) - y_i^{k-1} \right] \right]$

Matrix form $= x^{k-1} - \frac{\alpha}{n} \left[ \left(1 - \frac{1}{n}\right) e^T y^{k-1} + g'(x^{k-1}) + (z^k)^T \left[ f'(x^{k-1}) - y^{k-1} \right] \right]$

where $z^k = \begin{pmatrix} z_1^k I \\ \vdots \\ z_n^k I \end{pmatrix} \in \mathbb{R}^{np \times p}$

& we used the equality $g'(x) = \sum_{i=1}^{n} \frac{f_i'(x)}{n} = \frac{e^T f'(x)}{n}$

# Notations

- For $n$x$n$ matrix $M$, diag($M$): Vector of size $n$ composed of diagonal of $M$.

# Notations

- For $n$x$n$ matrix $M$, diag($M$): Vector of size $n$ composed of diagonal of $M$.
- For $n$-dimensional vector $m$, Diag($m$): $n$x$n$ diagonal matrix with m on it's diagonal.

# Notations

- For $nxn$ matrix $M$, diag($M$): Vector of size $n$ composed of diagonal of $M$.
- For $n$-dimensional vector $m$, Diag($m$): $nxn$ diagonal matrix with m on it's diagonal.
- $x^*$: unique minimizer of g. g is Strongly Convex so Strictly Convex also.

# Notations

- For $n$x$n$ matrix $M$, diag($M$): Vector of size $n$ composed of diagonal of $M$.
- For $n$-dimensional vector $m$, Diag($m$): $n$x$n$ diagonal matrix with m on it's diagonal.
- $x^*$: unique minimizer of g. g is Strongly Convex so Strictly Convex also.
- $\mathcal{F}_{k-1}$: $\sigma$-field of information generated by $z^1, z^2, ..., z^{k-1}$.

**To Prove**:
*With a constant step size of* $\alpha_k = \frac{1}{2nL}$*, the SAG iterations satisfy for* $k \geq 1$*:*

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leqslant \left(1 - \frac{\mu}{8Ln}\right)^k \left[3\|x_0 - x^*\|^2 + \frac{9\sigma^2}{4L^2}\right].$$

# Convergence Analysis

**To Prove**:
*With a constant step size of $\alpha_k = \frac{1}{2nL}$, the SAG iterations satisfy for $k \geq 1$:*

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leqslant \left(1 - \frac{\mu}{8Ln}\right)^k \left[3\|x_0 - x^*\|^2 + \frac{9\sigma^2}{4L^2}\right].$$

**Lyapunov functions** are scalar functions that can be used to verify stability of a dynamical system.

**To Prove**:
*With a constant step size of $\alpha_k = \frac{1}{2nL}$, the SAG iterations satisfy for $k \geq 1$:*

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leqslant \left(1 - \frac{\mu}{8Ln}\right)^k \left[3\|x_0 - x^*\|^2 + \frac{9\sigma^2}{4L^2}\right].$$

**Lyapunov functions** are scalar functions that can be used to verify stability of a dynamical system.
**Outline of Proof**:

# Convergence Analysis

**To Prove**:
*With a constant step size of $\alpha_k = \frac{1}{2nL}$, the SAG iterations satisfy for $k \geq 1$:*

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leqslant \left(1 - \frac{\mu}{8Ln}\right)^k \left[3\|x_0 - x^*\|^2 + \frac{9\sigma^2}{4L^2}\right].$$

**Lyapunov functions** are scalar functions that can be used to verify stability of a dynamical system.
**Outline of Proof**:

1. Find a Lyapunov function $Q$ from $\mathbb{R}^{(n+1)p}$ to $\mathbb{R}$ such that the sequence $\mathbb{E}[Q^k]$ decreases at a linear rate.

**To Prove**:
*With a constant step size of $\alpha_k = \frac{1}{2nL}$, the SAG iterations satisfy for $k \geq 1$:*

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leqslant \left(1 - \frac{\mu}{8Ln}\right)^k \left[3\|x_0 - x^*\|^2 + \frac{9\sigma^2}{4L^2}\right].$$

**Lyapunov functions** are scalar functions that can be used to verify stability of a dynamical system.
**Outline of Proof**:

1. Find a Lyapunov function $Q$ from $\mathbb{R}^{(n+1)p}$ to $\mathbb{R}$ such that the sequence $\mathbb{E}[Q^k]$ decreases at a linear rate.
2. Prove $Q(\theta^k)$ dominates $||x^k - x^*||^2$

**Step 1: Linear Convergence of Lyapunov Function** A quadratic $Q(\theta^k) = (\theta^k - \theta^*)^T P(\theta^k - \theta^*)$ is chosen as Lyapunov Function, where

$$\mathrm{P} = \begin{pmatrix} \mathrm{A} & \mathrm{b} \\ \mathrm{b}^T & \mathrm{c} \end{pmatrix}$$

**Step 1: Linear Convergence of Lyapunov Function** A quadratic $Q(\theta^k) = (\theta^k - \theta^*)^T P(\theta^k - \theta^*)$ is chosen as Lyapunov Function, where

$$P = \begin{pmatrix} A & b \\ b^T & c \end{pmatrix}$$

with

$$\begin{aligned} A &= 3n\alpha^2 I + \frac{\alpha^2}{n}(\frac{1}{n} - 2)ee^\top \\ b &= -\alpha(1 - \frac{1}{n})e \\ c &= I \\ S &= 3n\alpha^2 I \\ b - \frac{\alpha}{n}ec &= -\alpha e\ . \end{aligned}$$

**Step 1: Linear Convergence of Lyapunov Function** A quadratic $Q(\theta^k) = (\theta^k - \theta^*)^T P(\theta^k - \theta^*)$ is chosen as Lyapunov Function, where

$$P = \begin{pmatrix} A & b \\ b^T & c \end{pmatrix}$$

with

$$\begin{aligned} A &= 3n\alpha^2 I + \frac{\alpha^2}{n}(\frac{1}{n} - 2)ee^\top \\ b &= -\alpha(1 - \frac{1}{n})e \\ c &= I \\ S &= 3n\alpha^2 I \\ b - \frac{\alpha}{n}ec &= -\alpha e \; . \end{aligned}$$

**To show that:** $\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1})$ is negative for some $\delta > 0$.

# Convergence Analysis

**Expanding the first term**

$$\mathbb{E}\left[ (\theta^k - \theta^*)^\top \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix} (\theta^k - \theta^*) \middle| \mathcal{F}_{k-1} \right]$$
$$= E\left[ (y^k - f'(x^*))^\top A (y^k - f'(x^*)) + 2(y^k - f'(x^*))^\top b (x^k - x^*) + (x^k - x^*)^\top c (x^k - x^*) | \mathcal{F}_{k-1} \right]$$

**Expanding the first term**

$$\mathbb{E}\left[ (\theta^k - \theta^*)^\top \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix} (\theta^k - \theta^*) \middle| \mathcal{F}_{k-1} \right]$$
$$= E\left[ (y^k - f'(x^*))^\top A (y^k - f'(x^*)) + 2(y^k - f'(x^*))^\top b (x^k - x^*) + (x^k - x^*)^\top c (x^k - x^*) | \mathcal{F}_{k-1} \right]$$

Recall that:

- $y_i^k = (1 - \frac{1}{n}) y_i^{k-1} + \frac{1}{n} f_i'(x_i^{k-1}) + z_i^k [f_i'(x^{k-1}) - y_i^{k-1}]$

**Expanding the first term**

$$\mathbb{E}\left[ (\theta^k - \theta^*)^\top \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix} (\theta^k - \theta^*) \middle| \mathcal{F}_{k-1} \right]$$
$$= E\left[ (y^k - f'(x^*))^\top A (y^k - f'(x^*)) + 2(y^k - f'(x^*))^\top b(x^k - x^*) + (x^k - x^*)^\top c(x^k - x^*) | \mathcal{F}_{k-1} \right]$$

Recall that:

- $y_i^k = (1 - \frac{1}{n}) y_i^{k-1} + \frac{1}{n} f_i'(x_i^{k-1}) + z_i^k [f_i'(x^{k-1}) - y_i^{k-1}]$
- $x^k = x^{k-1} - \frac{\alpha}{n}[(1 - \frac{1}{n}) e^T y^{k-1} + g'(x^{k-1}) + (z^k)^T [f'(x^{k-1} - y^{k-1})]]$

**Expanding the first term**

$$\mathbb{E}\left[(\theta^k-\theta^*)^\top \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix} (\theta^k-\theta^*)\middle| \mathcal{F}_{k-1}\right]$$
$$= E\left[(y^k-f'(x^*))^\top A(y^k-f'(x^*))+2(y^k-f'(x^*))^\top b(x^k-x^*)+(x^k-x^*)^\top c(x^k-x^*)|\mathcal{F}_{k-1}\right]$$

Recall that:

- $y_i^k = (1-\frac{1}{n})y_i^{k-1} + \frac{1}{n}f_i'(x_i^{k-1}) + z_i^k[f_i'(x^{k-1}) - y_i^{k-1}]$
- $x^k = x^{k-1} - \frac{\alpha}{n}[(1-\frac{1}{n})e^T y^{k-1} + g'(x^{k-1}) + (z^k)^T[f'(x^{k-1} - y^{k-1})]]$
- $\mathcal{F}_{k-1}$: information upto $k-1$ iterations.

**Expanding the first term**

$$\mathbb{E}\left[ (\theta^k - \theta^*)^\top \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix} (\theta^k - \theta^*) \middle| \mathcal{F}_{k-1} \right]$$
$$= E\left[(y^k - f'(x^*))^\top A(y^k - f'(x^*)) + 2(y^k - f'(x^*))^\top b(x^k - x^*) + (x^k - x^*)^\top c(x^k - x^*)|\mathcal{F}_{k-1}\right]$$

Recall that:

- $y_i^k = (1 - \frac{1}{n})y_i^{k-1} + \frac{1}{n} f_i'(x_i^{k-1}) + z_i^k[f_i'(x^{k-1}) - y_i^{k-1}]$
- $x^k = x^{k-1} - \frac{\alpha}{n}[(1 - \frac{1}{n})e^T y^{k-1} + g'(x^{k-1}) + (z^k)^T[f'(x^{k-1} - y^{k-1})]]$
- $\mathcal{F}_{k-1}$: information upto $k-1$ iterations.
- $z_i^k$: Random variable that takes value $(1 - \frac{1}{n})$ with probability $\frac{1}{n}$ and take value $-\frac{1}{n}$ otherwise.

**Expanding the first term**

$$\mathbb{E}\left[(\theta^k-\theta^*)^\top \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix} (\theta^k-\theta^*)\middle| \mathcal{F}_{k-1}\right]$$
$$= E\left[(y^k - f'(x^*))^\top A(y^k - f'(x^*)) + 2(y^k - f'(x^*))^\top b(x^k - x^*) + (x^k - x^*)^\top c(x^k - x^*)|\mathcal{F}_{k-1}\right]$$

Recall that:

- $y_i^k = (1-\frac{1}{n})y_i^{k-1} + \frac{1}{n}f_i'(x_i^{k-1}) + z_i^k[f_i'(x^{k-1}) - y_i^{k-1}]$
- $x^k = x^{k-1} - \frac{\alpha}{n}[(1-\frac{1}{n})e^T y^{k-1} + g'(x^{k-1}) + (z^k)^T[f'(x^{k-1} - y^{k-1})]]$
- $\mathcal{F}_{k-1}$: information upto $k-1$ iterations.
- $z_i^k$: Random variable that takes value $(1-\frac{1}{n})$ with probability $\frac{1}{n}$ and take value $-\frac{1}{n}$ otherwise.
  - $\mathbb{E}[(z_i^k)^2] = (1-\frac{1}{n})^2(\frac{1}{n}) + (-\frac{1}{n})^2(1-\frac{1}{n}) = \frac{1}{n}(1-\frac{1}{n})$

**Expanding the first term**

$$\mathbb{E}\left[(\theta^k-\theta^*)^\top \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix} (\theta^k-\theta^*) \middle| \mathcal{F}_{k-1}\right]$$
$$= E\left[(y^k-f'(x^*))^\top A(y^k-f'(x^*)) + 2(y^k-f'(x^*))^\top b(x^k-x^*) + (x^k-x^*)^\top c(x^k-x^*)|\mathcal{F}_{k-1}\right]$$

Recall that:

- $y_i^k = (1-\frac{1}{n})y_i^{k-1} + \frac{1}{n}f_i'(x_i^{k-1}) + z_i^k[f_i'(x^{k-1}) - y_i^{k-1}]$
- $x^k = x^{k-1} - \frac{\alpha}{n}[(1-\frac{1}{n})e^T y^{k-1} + g'(x^{k-1}) + (z^k)^T[f'(x^{k-1} - y^{k-1})]]$
- $\mathcal{F}_{k-1}$: information upto $k-1$ iterations.
- $z_i^k$: Random variable that takes value $(1-\frac{1}{n})$ with probability $\frac{1}{n}$ and take value $-\frac{1}{n}$ otherwise.
  - $\mathbb{E}[(z_i^k)^2] = (1-\frac{1}{n})^2(\frac{1}{n}) + (-\frac{1}{n})^2(1-\frac{1}{n}) = \frac{1}{n}(1-\frac{1}{n})$
  - $\mathbb{E}[z_i^k z_j^k] = 2(1-\frac{1}{n})(\frac{-1}{n})(\frac{1}{n}) + (\frac{-1}{n})(\frac{-1}{n})\frac{n-2}{n} = -\frac{1}{n^2}$

**Expanding the first term**

$$\mathbb{E}\left[(\theta^k-\theta^*)^\top \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix} (\theta^k-\theta^*) \middle| \mathcal{F}_{k-1}\right]$$
$$= E\left[(y^k - f'(x^*))^\top A(y^k - f'(x^*)) + 2(y^k - f'(x^*))^\top b(x^k - x^*) + (x^k - x^*)^\top c(x^k - x^*)|\mathcal{F}_{k-1}\right]$$

Recall that:

- $y_i^k = (1-\frac{1}{n})y_i^{k-1} + \frac{1}{n}f_i'(x_i^{k-1}) + z_i^k[f_i'(x^{k-1}) - y_i^{k-1}]$
- $x^k = x^{k-1} - \frac{\alpha}{n}[(1-\frac{1}{n})e^T y^{k-1} + g'(x^{k-1}) + (z^k)^T[f'(x^{k-1} - y^{k-1})]]$
- $\mathcal{F}_{k-1}$: information upto $k-1$ iterations.
- $z_i^k$: Random variable that takes value $(1-\frac{1}{n})$ with probability $\frac{1}{n}$ and take value $-\frac{1}{n}$ otherwise.
  - $\mathbb{E}[(z_i^k)^2] = (1-\frac{1}{n})^2(\frac{1}{n}) + (-\frac{1}{n})^2(1-\frac{1}{n}) = \frac{1}{n}(1-\frac{1}{n})$
  - $\mathbb{E}[z_i^k z_j^k] = 2(1-\frac{1}{n})(\frac{-1}{n})(\frac{1}{n}) + (\frac{-1}{n})(\frac{-1}{n})\frac{n-2}{n} = -\frac{1}{n^2}$

The randomness in $Q(\theta^k)$ comes only because of $z^k$

# Convergence Analysis

After substituting the expectation of $z^k$ & re-arranging,
$\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}]$

$$
\begin{aligned}
&= \left(1 - \frac{1}{n}\right) 3n\alpha^2 (y^{k-1} - f'(x^*))^\top (y^{k-1} - f'(x^*)) \\
&\quad + (x^{k-1} - x^*)^\top (x^{k-1} - x^*) - 2\alpha (x^{k-1} - x^*)^\top g'(x^{k-1}) \\
&\quad + 3\alpha^2 (f'(x^{k-1}) - f'(x^*))^\top (f'(x^{k-1}) - f'(x^*)) \\
&\quad - 2\alpha \left(1 - \frac{1}{n}\right) (y^{k-1} - f'(x^*))^\top e(x^{k-1} - x^*)
\end{aligned}
$$

After substituting the expectation of $z^k$ & re-arranging,
$\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}]$

$$
\begin{aligned}
&= \left(1-\frac{1}{n}\right) 3n\alpha^2 (y^{k-1} - f'(x^*))^\top (y^{k-1} - f'(x^*)) \\
&\quad + (x^{k-1} - x^*)^\top (x^{k-1} - x^*) - 2\alpha (x^{k-1} - x^*)^\top g'(x^{k-1}) \\
&\quad + 3\alpha^2 (f'(x^{k-1}) - f'(x^*))^\top (f'(x^{k-1}) - f'(x^*)) \\
&\quad - 2\alpha \left(1-\frac{1}{n}\right) (y^{k-1} - f'(x^*))^\top e(x^{k-1} - x^*)
\end{aligned}
$$

- The goal is to bound $\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1})$ by $g'(x^{k-1})^T(x^{k-1} - x^*)$ positive due to Convexity.

After substituting the expectation of $z^k$ & re-arranging,
$\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}]$

$$
\begin{aligned}
&= \left(1 - \frac{1}{n}\right) 3n\alpha^2 (y^{k-1} - f'(x^*))^\top (y^{k-1} - f'(x^*)) \\
&\quad + (x^{k-1} - x^*)^\top (x^{k-1} - x^*) - 2\alpha (x^{k-1} - x^*)^\top g'(x^{k-1}) \\
&\quad + 3\alpha^2 (f'(x^{k-1}) - f'(x^*))^\top (f'(x^{k-1}) - f'(x^*)) \\
&\quad - 2\alpha \left(1 - \frac{1}{n}\right) (y^{k-1} - f'(x^*))^\top e(x^{k-1} - x^*)
\end{aligned}
$$

- The goal is to bound $\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1})$ by $g'(x^{k-1})^T(x^{k-1} - x^*)$ positive due to Convexity.
- The $3^{rd}$ term contains $(f'(x^{k-1}) - f'(x^*))^T(f'(x^{k-1}) - f'(x^*)) = \sum_{i=1}^{n} ||(f_i'(x^{k-1}) - f_i'(x^*))||^2$

After substituting the expectation of $z^k$ & re-arranging,
$\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}]$

$$= \left(1-\frac{1}{n}\right)3n\alpha^2(y^{k-1}-f'(x^*))^\top(y^{k-1}-f'(x^*))$$
$$+(x^{k-1}-x^*)^\top(x^{k-1}-x^*)-2\alpha(x^{k-1}-x^*)^\top g'(x^{k-1})$$
$$+3\alpha^2(f'(x^{k-1})-f'(x^*))^\top(f'(x^{k-1})-f'(x^*))$$
$$-2\alpha\left(1-\frac{1}{n}\right)(y^{k-1}-f'(x^*))^\top e(x^{k-1}-x^*)$$

- The goal is to bound $\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}]-(1-\delta)Q(\theta^{k-1})$ by $g'(x^{k-1})^T(x^{k-1}-x^*)$ positive due to Convexity.
- The $3^{rd}$ term contains $(f'(x^{k-1})-f'(x^*))^T(f'(x^{k-1})-f'(x^*)) = \sum_{i=1}^n ||(f_i'(x^{k-1})-f_i'(x^*))||^2$
$\leq \sum_{i=1}^n (f_i'(x^{k-1})-f_i'(x^*))^T L(x^{k-1}-x^*) = nLg'(x^{k-1})^T(x^{k-1}-x^*)$
gradients' L-Lipschitz continuity & g'(x*)=0.

$$\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \leq (y^{k-1} - f'(x^*))^T[3n\alpha^2(\delta - \tfrac{1}{n})I + (1-\delta)\tfrac{\alpha^2}{n}(2 - \tfrac{1}{n})ee^T](y^{k-1} - f'(x^*)) + \delta(x^{k-1} - x^*)^T(x^{k-1} - x^*) - (2\alpha - 3\alpha^2 nL)(x^{k-1} - x^*)^T g'(x^{k-1}) + (y^{k-1} - f'(x^*))^T(-2\alpha(1 - \tfrac{1}{n})e)(x^{k-1} - x^*)$$

$$\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \leq (y^{k-1} - f'(x^*))^T[3n\alpha^2(\delta - \frac{1}{n})I + (1-\delta)\frac{\alpha^2}{n}(2-\frac{1}{n})ee^T](y^{k-1} - f'(x^*)) + \delta(x^{k-1}-x^*)^T(x^{k-1}-x^*) - (2\alpha - 3\alpha^2 nL)(x^{k-1} - x^*)^T g'(x^{k-1}) + (y^{k-1} - f'(x^*))^T(-2\alpha(1-\frac{1}{n})e)(x^{k-1}-x^*)$$

- The RHS contains terms of the form: $s^T Ms + s^T t \leq -\frac{1}{4}t^T M^{-1} t$ when M is negative-definite.

$$\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \leq (y^{k-1} - f'(x^*))^T[3n\alpha^2(\delta - \tfrac{1}{n})I + (1-\delta)\tfrac{\alpha^2}{n}(2-\tfrac{1}{n})ee^T](y^{k-1} - f'(x^*)) + \delta(x^{k-1}-x^*)^T(x^{k-1}-x^*) - (2\alpha - 3\alpha^2 nL)(x^{k-1} - x^*)^T g'(x^{k-1}) + (y^{k-1} - f'(x^*))^T(-2\alpha(1-\tfrac{1}{n})e)(x^{k-1}-x^*)$$

- The RHS contains terms of the form: $s^T M s + s^T t \leq -\frac{1}{4} t^T M^{-1} t$ when M is negative-definite.
  - For $M \prec 0$, $s^T M s + s^T t = -\lambda||s||^2 + s^T t + (-\frac{1}{4\lambda}||t||^2) = -||2\lambda s - t||^2 \leq 0$

$$\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \leq (y^{k-1} - f'(x^*))^T[3n\alpha^2(\delta - \tfrac{1}{n})I + (1-\delta)\tfrac{\alpha^2}{n}(2 - \tfrac{1}{n})ee^T](y^{k-1} - f'(x^*)) + \delta(x^{k-1} - x^*)^T(x^{k-1} - x^*) - (2\alpha - 3\alpha^2 nL)(x^{k-1} - x^*)^T g'(x^{k-1}) + (y^{k-1} - f'(x^*))^T(-2\alpha(1 - \tfrac{1}{n})e)(x^{k-1} - x^*)$$

- The RHS contains terms of the form: $s^T M s + s^T t \leq -\frac{1}{4} t^T M^{-1} t$ when M is negative-definite.
  - For $M \prec 0$, $s^T M s + s^T t = -\lambda||s||^2 + s^T t + (-\frac{1}{4\lambda}||t||^2) = -||2\lambda s - t||^2 \leq 0$
  - Sufficient condition for $M \prec 0$ is $\delta \leq \frac{1}{3n}$

Using this,

$$\begin{aligned}\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \leqslant {} & -(2\alpha - 3\alpha^2 nL)(x^{k-1} - x^*)^\top g'(x^{k-1}) \\ & + \left(\delta - \frac{\delta^2\left(1-\frac{1}{n}\right)^2}{\left[3n\delta - 1 - 2\delta + \frac{\delta-1}{n}\right]} n\right)\|x^{k-1} - x^*\|^2\end{aligned}$$

Using this,

$$\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \leqslant -(2\alpha - 3\alpha^2 nL)(x^{k-1} - x^*)^\top g'(x^{k-1}) + \left(\delta - \frac{\delta^2\left(1-\frac{1}{n}\right)^2}{\left[3n\delta - 1 - 2\delta + \frac{\delta-1}{n}\right]} n\right) \|x^{k-1} - x^*\|^2$$

- Using $||x^{k-1} - x^*||^2 \leq \frac{1}{\mu}(x^{k-1} - x^*)^T g'(x^{k-1})$ Strong Convexity.
  $\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \leq$
  $$-\left(2\alpha - 3\alpha^2 nL + \frac{\delta^2\left(1-\frac{1}{n}\right)^2}{\left[3n\delta - 1 - 2\delta + \frac{\delta-1}{n}\right]}\frac{n}{\mu} - \frac{\delta}{\mu}\right)(x^{k-1} - x^*)^\top g'(x^{k-1})$$

Using this,

$$\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \leqslant -(2\alpha - 3\alpha^2 nL)(x^{k-1}-x^*)^\top g'(x^{k-1}) + \left(\delta - \frac{\delta^2\left(1-\frac{1}{n}\right)^2}{\left[3n\delta - 1 - 2\delta + \frac{\delta-1}{n}\right]}n\right)\|x^{k-1}-x^*\|^2$$

- Using $||x^{k-1}-x^*||^2 \leq \frac{1}{\mu}(x^{k-1}-x^*)^T g'(x^{k-1})$ Strong Convexity.
  $\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \leq$
  $$-\left(2\alpha - 3\alpha^2 nL + \frac{\delta^2\left(1-\frac{1}{n}\right)^2}{\left[3n\delta - 1 - 2\delta + \frac{\delta-1}{n}\right]}\frac{n}{\mu} - \frac{\delta}{\mu}\right)(x^{k-1}-x^*)^\top g'(x^{k-1})$$
- $(x^{k-1}-x^*)^T g'(x^{k-1})$ is positive due to Convexity . The coefficient(without minus) is positive for step size $\alpha = \frac{1}{2nL}$ with $\delta = \frac{\mu}{8nL}$.

Using this,

$$\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \leqslant -(2\alpha - 3\alpha^2 nL)(x^{k-1}-x^*)^\top g'(x^{k-1}) + \left(\delta - \frac{\delta^2\left(1-\frac{1}{n}\right)^2}{\left[3n\delta - 1 - 2\delta + \frac{\delta-1}{n}\right]} n\right)\|x^{k-1}-x^*\|^2$$

- Using $||x^{k-1} - x^*||^2 \leq \frac{1}{\mu}(x^{k-1} - x^*)^T g'(x^{k-1})$ Strong Convexity.
  $\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \leq$
  $$-\left(2\alpha - 3\alpha^2 nL + \frac{\delta^2\left(1-\frac{1}{n}\right)^2}{\left[3n\delta - 1 - 2\delta + \frac{\delta-1}{n}\right]}\frac{n}{\mu} - \frac{\delta}{\mu}\right)(x^{k-1}-x^*)^\top g'(x^{k-1})$$
- $(x^{k-1} - x^*)^T g'(x^{k-1})$ is positive due to Convexity . The coefficient(without minus) is positive for step size $\alpha = \frac{1}{2nL}$ with $\delta = \frac{\mu}{8nL}$.
- Using Total Expectation, $\mathbb{E}[Q(\theta^k)] - (1 - \frac{\mu}{8nL})\mathbb{E}[Q(\theta^{k-1})] \leq 0$

Using this,

$$\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \leqslant -(2\alpha - 3\alpha^2 nL)(x^{k-1} - x^*)^\top g'(x^{k-1}) + \left(\delta - \frac{\delta^2\left(1-\frac{1}{n}\right)^2}{\left[3n\delta - 1 - 2\delta + \frac{\delta-1}{n}\right]} n\right) \|x^{k-1} - x^*\|^2$$

- Using $||x^{k-1} - x^*||^2 \leq \frac{1}{\mu}(x^{k-1} - x^*)^T g'(x^{k-1})$ Strong Convexity.
  $\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \leq$
  $$-\left(2\alpha - 3\alpha^2 nL + \frac{\delta^2\left(1-\frac{1}{n}\right)^2}{\left[3n\delta - 1 - 2\delta + \frac{\delta-1}{n}\right]}\frac{n}{\mu} - \frac{\delta}{\mu}\right)(x^{k-1} - x^*)^\top g'(x^{k-1})$$
- $(x^{k-1} - x^*)^T g'(x^{k-1})$ is positive due to Convexity . The coefficient(without minus) is positive for step size $\alpha = \frac{1}{2nL}$ with $\delta = \frac{\mu}{8nL}$.
- Using Total Expectation, $\mathbb{E}[Q(\theta^k)] - (1 - \frac{\mu}{8nL})\mathbb{E}[Q(\theta^{k-1})] \leq 0$

*Step 1(Linear Convergence of Lyapunov Function) finished.*

# Convergence Analysis

**Step 2:** $Q(\theta^k)$ dominates $||x^k - x^*||^2$

**Step 2:** $Q(\theta^k)$ dominates $||x^k - x^*||^2$

- Recall that $Q(\theta^k) = (\theta^k - \theta^*)^T P(\theta^k - \theta^*)$, where $\mathrm{P} = \begin{pmatrix} \mathrm{A} & \mathrm{b} \\ \mathrm{b}^T & \mathrm{c} \end{pmatrix}$; $A$ was positive definite.

# Convergence Analysis

**Step 2:** $Q(\theta^k)$ dominates $||x^k - x^*||^2$

- Recall that $Q(\theta^k) = (\theta^k - \theta^*)^T P(\theta^k - \theta^*)$, where $\mathrm{P} = \begin{pmatrix} \mathrm{A} & \mathrm{b} \\ \mathrm{b}^T & \mathrm{c} \end{pmatrix}$; $A$ was positive definite.
- Analysing $(\theta^k - \theta^*)^T (P - \begin{pmatrix} 0 & 0 \\ 0 & \frac{I}{3} \end{pmatrix})(\theta^k - \theta^*)$

**Step 2:** $Q(\theta^k)$ dominates $||x^k - x^*||^2$

- Recall that $Q(\theta^k) = (\theta^k - \theta^*)^T P(\theta^k - \theta^*)$, where $\mathrm{P} = \begin{pmatrix} \mathrm{A} & \mathrm{b} \\ \mathrm{b}^T & \mathrm{c} \end{pmatrix}$; $A$ was positive definite.
- Analysing $(\theta^k - \theta^*)^T (P - \begin{pmatrix} 0 & 0 \\ 0 & \frac{I}{3} \end{pmatrix})(\theta^k - \theta^*)$
- $(P - \begin{pmatrix} 0 & 0 \\ 0 & \frac{I}{3} \end{pmatrix}) = \begin{pmatrix} \mathrm{A} & \mathrm{b} \\ \mathrm{b}^T & c - \frac{I}{3} \end{pmatrix} \succ 0$ iff Schur complement of the block $A = (c - \frac{I}{3}) - b^T A^{-1} b \succ 0$. Schur Complement Lemma

# Convergence Analysis

**Step 2:** $Q(\theta^k)$ dominates $||x^k - x^*||^2$

- Recall that $Q(\theta^k) = (\theta^k - \theta^*)^T P(\theta^k - \theta^*)$, where $\mathrm{P} = \begin{pmatrix} \mathrm{A} & \mathrm{b} \\ \mathrm{b}^T & \mathrm{c} \end{pmatrix}$; $A$ was positive definite.
- Analysing $(\theta^k - \theta^*)^T(P - \begin{pmatrix} 0 & 0 \\ 0 & \frac{I}{3} \end{pmatrix})(\theta^k - \theta^*)$
- $(P - \begin{pmatrix} 0 & 0 \\ 0 & \frac{I}{3} \end{pmatrix}) = \begin{pmatrix} \mathrm{A} & \mathrm{b} \\ \mathrm{b}^T & c - \frac{I}{3} \end{pmatrix} \succ 0$ iff Schur complement of the block $A = (c - \frac{I}{3}) - b^T A^{-1} b \succ 0$. Schur Complement Lemma
- For $N \geq 2, (P - \begin{pmatrix} 0 & 0 \\ 0 & \frac{I}{3} \end{pmatrix}) \succ 0 \implies Q(\theta^k) \geq \frac{1}{3}||x^k - x^*||^2$

**Step 2:** $Q(\theta^k)$ dominates $||x^k - x^*||^2$

- Recall that $Q(\theta^k) = (\theta^k - \theta^*)^T P(\theta^k - \theta^*)$, where $\mathrm{P} = \begin{pmatrix} \mathrm{A} & \mathrm{b} \\ \mathrm{b}^T & \mathrm{c} \end{pmatrix}$; $A$ was positive definite.
- Analysing $(\theta^k - \theta^*)^T (P - \begin{pmatrix} 0 & 0 \\ 0 & \frac{I}{3} \end{pmatrix})(\theta^k - \theta^*)$
- $(P - \begin{pmatrix} 0 & 0 \\ 0 & \frac{I}{3} \end{pmatrix}) = \begin{pmatrix} \mathrm{A} & \mathrm{b} \\ \mathrm{b}^T & c - \frac{I}{3} \end{pmatrix} \succ 0$ iff Schur complement of the block $A = (c - \frac{I}{3}) - b^T A^{-1} b \succ 0$. Schur Complement Lemma
- For $N \geq 2, (P - \begin{pmatrix} 0 & 0 \\ 0 & \frac{I}{3} \end{pmatrix}) \succ 0 \implies Q(\theta^k) \geq \frac{1}{3}||x^k - x^*||^2$
- $\mathbb{E}||x^k - x^*||^2 \leq 3\mathbb{E}[\theta^k]$. With all $y_i^0 = 0$, $\mathbb{E}||x^k - x^*||^2 \leq (1 - \frac{\mu}{8NL})^k (\frac{9\sigma^2}{4L^2} + 3||x^k - x^*||^2)$

# Convergence Analysis

- Theorem for faster rate: With $\alpha_t = \frac{1}{16L}$ SAG iteration satisfy $\mathbb{E}[g(x^k) - g(x^*)] \leq (1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^k C$

# Convergence Analysis

- Theorem for faster rate: With $\alpha_t = \frac{1}{16L}$ SAG iteration satisfy $\mathbb{E}[g(x^k) - g(x^*)] \leq (1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^k C$
- For well-conditioned problem, constant reduction per pass.

# Convergence Analysis

- Theorem for faster rate: With $\alpha_t = \frac{1}{16L}$ SAG iteration satisfy $\mathbb{E}[g(x^k) - g(x^*)] \leq (1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^k C$
- For well-conditioned problem, constant reduction per pass.
- Rate Comparison with Deterministic methods:
  - Gradient Method(Step-size $\frac{2}{\mu+L}$) $(\frac{L-\mu}{L+\mu})^2$

# Convergence Analysis

- Theorem for faster rate: With $\alpha_t = \frac{1}{16L}$ SAG iteration satisfy $\mathbb{E}[g(x^k) - g(x^*)] \leq (1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^k C$
- For well-conditioned problem, constant reduction per pass.
- Rate Comparison with Deterministic methods:
  - Gradient Method(Step-size $\frac{2}{\mu+L}$) $(\frac{L-\mu}{L+\mu})^2$
  - Accelerated Gradient method(Step-size $\frac{1}{L}$) $(1 - \sqrt{\frac{\mu}{L}})$

# Convergence Analysis

- Theorem for faster rate: With $\alpha_t = \frac{1}{16L}$ SAG iteration satisfy $\mathbb{E}[g(x^k) - g(x^*)] \leq (1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^k C$
- For well-conditioned problem, constant reduction per pass.
- Rate Comparison with Deterministic methods:
  - Gradient Method(Step-size $\frac{2}{\mu+L}$) $(\frac{L-\mu}{L+\mu})^2$
  - Accelerated Gradient method(Step-size $\frac{1}{L}$) $(1 - \sqrt{\frac{\mu}{L}})$
  - Fastest possible first-order method $(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}})^2$

# Convergence Analysis

- Theorem for faster rate: With $\alpha_t = \frac{1}{16L}$ SAG iteration satisfy $\mathbb{E}[g(x^k) - g(x^*)] \leq (1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^k C$
- For well-conditioned problem, constant reduction per pass.
- Rate Comparison with Deterministic methods:
  - Gradient Method(Step-size $\frac{2}{\mu+L}$) $(\frac{L-\mu}{L+\mu})^2$
  - Accelerated Gradient method(Step-size $\frac{1}{L}$) $(1 - \sqrt{\frac{\mu}{L}})$
  - Fastest possible first-order method $(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}})^2$
  - SAG(N iterations)(Step-size $\frac{1}{16L}$) $(1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^N$

# Convergence Analysis

- Theorem for faster rate: With $\alpha_t = \frac{1}{16L}$ SAG iteration satisfy $\mathbb{E}[g(x^k) - g(x^*)] \leq (1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^k C$
- For well-conditioned problem, constant reduction per pass.
- Rate Comparison with Deterministic methods:
  - Gradient Method(Step-size $\frac{2}{\mu+L}$) $(\frac{L-\mu}{L+\mu})^2$
  - Accelerated Gradient method(Step-size $\frac{1}{L}$) $(1 - \sqrt{\frac{\mu}{L}})$
  - Fastest possible first-order method $(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}})^2$
  - SAG(N iterations)(Step-size $\frac{1}{16L}$) $(1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^N$
- No. of $f_i'$ evaluations to reach $\epsilon$
  - Stochastic: $O(\frac{L}{\mu}(\frac{1}{\epsilon}))$

# Convergence Analysis

- Theorem for faster rate: With $\alpha_t = \frac{1}{16L}$ SAG iteration satisfy $\mathbb{E}[g(x^k) - g(x^*)] \leq (1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^k C$
- For well-conditioned problem, constant reduction per pass.
- Rate Comparison with Deterministic methods:
  - Gradient Method(Step-size $\frac{2}{\mu+L}$) $(\frac{L-\mu}{L+\mu})^2$
  - Accelerated Gradient method(Step-size $\frac{1}{L}$) $(1 - \sqrt{\frac{\mu}{L}})$
  - Fastest possible first-order method $(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}})^2$
  - SAG(N iterations)(Step-size $\frac{1}{16L}$) $(1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^N$
- No. of $f_i'$ evaluations to reach $\epsilon$
  - Stochastic: $O(\frac{L}{\mu}(\frac{1}{\epsilon}))$
  - Gradient: $O(N\frac{L}{\mu}(\frac{1}{\epsilon}))$

- Theorem for faster rate: With $\alpha_t = \frac{1}{16L}$ SAG iteration satisfy $\mathbb{E}[g(x^k) - g(x^*)] \leq (1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^k C$
- For well-conditioned problem, constant reduction per pass.
- Rate Comparison with Deterministic methods:
  - Gradient Method(Step-size $\frac{2}{\mu+L}$) $(\frac{L-\mu}{L+\mu})^2$
  - Accelerated Gradient method(Step-size $\frac{1}{L}$) $(1 - \sqrt{\frac{\mu}{L}})$
  - Fastest possible first-order method $(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}})^2$
  - SAG(N iterations)(Step-size $\frac{1}{16L}$) $(1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^N$
- No. of $f_i'$ evaluations to reach $\epsilon$
  - Stochastic: $O(\frac{L}{\mu}(\frac{1}{\epsilon}))$
  - Gradient: $O(N\frac{L}{\mu}(\frac{1}{\epsilon}))$
  - Accelerated: $O(\sqrt{\frac{L}{\mu}}(\frac{1}{\epsilon}))$

# Convergence Analysis

- Theorem for faster rate: With $\alpha_t = \frac{1}{16L}$ SAG iteration satisfy $\mathbb{E}[g(x^k) - g(x^*)] \leq (1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^k C$
- For well-conditioned problem, constant reduction per pass.
- Rate Comparison with Deterministic methods:
  - Gradient Method(Step-size $\frac{2}{\mu+L}$) $(\frac{L-\mu}{L+\mu})^2$
  - Accelerated Gradient method(Step-size $\frac{1}{L}$) $(1 - \sqrt{\frac{\mu}{L}})$
  - Fastest possible first-order method $(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}})^2$
  - SAG(N iterations)(Step-size $\frac{1}{16L}$) $(1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^N$
- No. of $f_i'$ evaluations to reach $\epsilon$
  - Stochastic: $O(\frac{L}{\mu}(\frac{1}{\epsilon}))$
  - Gradient: $O(N\frac{L}{\mu}(\frac{1}{\epsilon}))$
  - Accelerated: $O(\sqrt{\frac{L}{\mu}}(\frac{1}{\epsilon}))$
  - **SAG**: $O(\max\{N, \frac{L}{\mu}\}log(\frac{1}{\epsilon}))$

# Implementation Issues

**Algorithm 1** Basic SAG method for minimizing $\frac{1}{n}\sum_{i=1}^{n} f_i(x)$ with step size $\alpha$

$d = 0$, $y_i = 0$ for $i = 1, 2, \ldots, n$
**for** $k = 0, 1, \ldots$ **do**
  Sample $i$ from $\{1, 2, \ldots, n\}$
  $d = d - y_i + f_i'(x)$
  $y_i = f_i'(x)$
  $x = x - \frac{\alpha}{n} d$
**end for**

# Implementation Issues

---

**Algorithm 1** Basic SAG method for minimizing $\frac{1}{n}\sum_{i=1}^{n} f_i(x)$ with step size $\alpha$

---

$d = 0$, $y_i = 0$ for $i = 1, 2, \ldots, n$
**for** $k = 0, 1, \ldots$ **do**
  Sample $i$ from $\{1, 2, \ldots, n\}$
  $d = d - y_i + f_i'(x)$
  $y_i = f_i'(x)$
  $x = x - \frac{\alpha}{n} d$
**end for**

---

Termination criteria: When $||\frac{d}{N}|| = ||\frac{\sum_{i=1}^{N} y_i}{N}||$ becomes small.

1. **Structured Gradients**: For linearly parameterized models $\min_{x \in \mathbb{R}^P} g(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(a_i^T x)$. Only store $f'_{i_k}(a_{i_k}^T x^k)$. Storage cost reduces from $O(NP)$ to $O(N)$.

# Implementation Issues

1. **Structured Gradients**: For linearly parameterized models $\min_{x \in \mathbb{R}^P} g(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(a_i^T x)$. Only store $f'_{i_k}(a_{i_k}^T x^k)$. Storage cost reduces from $O(NP)$ to $O(N)$.
   - **Just-in-time updates**: If vectors $a_i$ are sparse: Instead of storing $x^k$ after each iterataion, only compute $x_j^k$ corresponding to non-zero elements of $a_{i_k}$. Iteration cost reduces from $O(P)$ to $O(||f_i'(x)||_0)$.

# Implementation Issues

1. **Structured Gradients**: For linearly parameterized models $\min_{x \in \mathbb{R}^P} g(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(a_i^T x)$. Only store $f'_{i_k}(a_{i_k}^T x^k)$. Storage cost reduces from O($NP$) to O($N$).
   - **Just-in-time updates**: If vectors $a_i$ are sparse: Instead of storing $x^k$ after each iterataion, only compute $x_j^k$ corresponding to non-zero elements of $a_{i_k}$. Iteration cost reduces from O($P$) to O($||f_i'(x)||_0$).
2. **Normalization** $x = x - \frac{\alpha}{m} d$ where m is the no. of data points seen at least once.

# Implementation Issues

1. **Structured Gradients**: For linearly parameterized models $\min_{x \in \mathbb{R}^P} g(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(a_i^T x)$. Only store $f_{i_k}'(a_{i_k}^T x^k)$. Storage cost reduces from $O(NP)$ to $O(N)$.
   - **Just-in-time updates**: If vectors $a_i$ are sparse: Instead of storing $x^k$ after each iterataion, only compute $x_j^k$ corresponding to non-zero elements of $a_{i_k}$. Iteration cost reduces from $O(P)$ to $O(||f_i'(x)||_0)$.
2. **Normalization** $x = x - \frac{\alpha}{m} d$ where m is the no. of data points seen at least once.
3. **Step-size** $\frac{1}{L}$ in practice. When $L$ is not known:
   - Start with a small $L$.
   - Double $L$ if $f_{i_k}(x^k - \frac{1}{L^k} f_{i_k}'(x^k)) \leq f_{i_k}(x^k) - \frac{1}{2L^k} ||f_{i_k}'(x^k)||^2$ not satisfied. Smoothness

# Implementation Issues

1. **Structured Gradients**: For linearly parameterized models $\min_{x \in \mathbb{R}^P} g(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(a_i^T x)$. Only store $f'_{i_k}(a_{i_k}^T x^k)$. Storage cost reduces from $O(NP)$ to $O(N)$.
   - **Just-in-time updates**: If vectors $a_i$ are sparse: Instead of storing $x^k$ after each iterataion, only compute $x_j^k$ corresponding to non-zero elements of $a_{i_k}$. Iteration cost reduces from $O(P)$ to $O(||f_i'(x)||_0)$.
2. **Normalization** $x = x - \frac{\alpha}{m} d$ where m is the no. of data points seen at least once.
3. **Step-size** $\frac{1}{L}$ in practice. When $L$ is not known:
   - Start with a small $L$.
   - Double $L$ if $f_{i_k}(x^k - \frac{1}{L^k} f'_{i_k}(x^k)) \leq f_{i_k}(x^k) - \frac{1}{2L^k} ||f'_{i_k}(x^k)||^2$ not satisfied.Smoothness
4. **Mini-batches** Batch-size of $k$ leads to $k$-fold reduction in storage cost. L can be chosen to be one of max eigenvalue$\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} L_i \leq \text{mean}_{i \in \mathcal{B}}\{L_i\} \leq \max_{i \in \mathcal{B}}\{L_i\}$

- **Non-uniform example selection**: Convergence rate depends on $\frac{\mu}{L}$. One way to improve convergence rate by decreasing $L$.(For a sum of functions $L$ is usually the max of all Lipschitz constants)
  - We can replace $f_n$ by $f_{n1}$ & $f_{n2}$ st. $f_{n1}(x) = f_{n2}(x) = \frac{f_n(x)}{2}$, making $g(x) = \frac{1}{n}(\sum_{i=1}^{n-1} f_i(x) + f_{n1}(x) + f_{n2}(x))$ which can be written as $g(x) = \frac{1}{n+1}(\sum_{i=1}^{n-1} \frac{n+1}{n} f_i(x) + \frac{n+1}{n} f_{n1}(x) + \frac{n+1}{n} f_{n2}(x))$

- **Non-uniform example selection**: Convergence rate depends on $\frac{\mu}{L}$. One way to improve convergence rate by decreasing $L$.(For a sum of functions $L$ is usually the max of all Lipschitz constants)
  - We can replace $f_n$ by $f_{n1}$ & $f_{n2}$ st. $f_{n1}(x) = f_{n2}(x) = \frac{f_n(x)}{2}$, making $g(x) = \frac{1}{n}(\sum_{i=1}^{n-1} f_i(x) + f_{n1}(x) + f_{n2}(x))$ which can be written as $g(x) = \frac{1}{n+1}(\sum_{i=1}^{n-1} \frac{n+1}{n} f_i(x) + \frac{n+1}{n} f_{n1}(x) + \frac{n+1}{n} f_{n2}(x))$
    - $\frac{n+1}{n}|f_{n1}(x) - f_{n1}(y)| = \frac{n+1}{n}\frac{|f_n(x)-f_n(y)|}{2} \leq \frac{n+1}{2n} L_n||x-y||^2$

# Implementation Issues

- **Non-uniform example selection**: Convergence rate depends on $\frac{\mu}{L}$. One way to improve convergence rate by decreasing $L$.(For a sum of functions $L$ is usually the max of all Lipschitz constants)
  - We can replace $f_n$ by $f_{n1}$ & $f_{n2}$ st. $f_{n1}(x) = f_{n2}(x) = \frac{f_n(x)}{2}$, making $g(x) = \frac{1}{n}(\sum_{i=1}^{n-1} f_i(x) + f_{n1}(x) + f_{n2}(x))$ which can be written as $g(x) = \frac{1}{n+1}(\sum_{i=1}^{n-1} \frac{n+1}{n} f_i(x) + \frac{n+1}{n} f_{n1}(x) + \frac{n+1}{n} f_{n2}(x))$
    - $\frac{n+1}{n}|f_{n1}(x) - f_{n1}(y)| = \frac{n+1}{n}\frac{|f_n(x) - f_n(y)|}{2} \leq \frac{n+1}{2n} L_n||x-y||^2$
    - Also increased probability of $f_n$ being sampled from $\frac{1}{n}$ to $\frac{2}{(n+1)}$

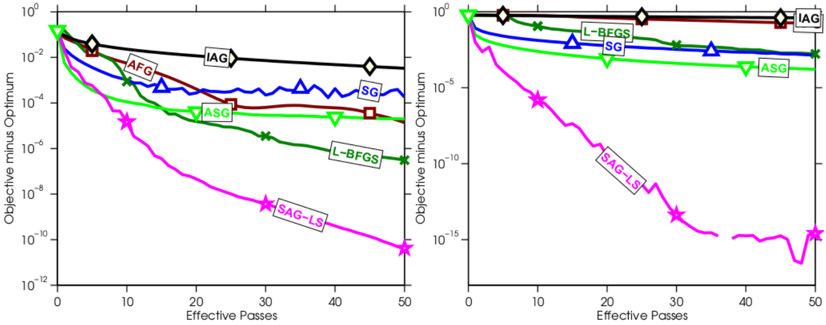
# Implementation Issues

- **Non-uniform example selection**: Convergence rate depends on $\frac{\mu}{L}$. One way to improve convergence rate by decreasing $L$.(For a sum of functions $L$ is usually the max of all Lipschitz constants)
  - We can replace $f_n$ by $f_{n1}$ & $f_{n2}$ st. $f_{n1}(x) = f_{n2}(x) = \frac{f_n(x)}{2}$, making $g(x) = \frac{1}{n}(\sum_{i=1}^{n-1} f_i(x) + f_{n1}(x) + f_{n2}(x))$ which can be written as $g(x) = \frac{1}{n+1}(\sum_{i=1}^{n-1} \frac{n+1}{n} f_i(x) + \frac{n+1}{n} f_{n1}(x) + \frac{n+1}{n} f_{n2}(x))$
    - $\frac{n+1}{n}|f_{n1}(x) - f_{n1}(y)| = \frac{n+1}{n}\frac{|f_n(x) - f_n(y)|}{2} \leq \frac{n+1}{2n} L_n ||x-y||^2$
    - Also increased probability of $f_n$ being sampled from $\frac{1}{n}$ to $\frac{2}{(n+1)}$
  - Duplicate each function $f_i$ a no. of times equal to it's Lipschitz constant of their gradient.

- **Non-uniform example selection**: Convergence rate depends on $\frac{\mu}{L}$. One way to improve convergence rate by decreasing $L$.(For a sum of functions $L$ is usually the max of all Lipschitz constants)
  - We can replace $f_n$ by $f_{n1}$ & $f_{n2}$ st. $f_{n1}(x) = f_{n2}(x) = \frac{f_n(x)}{2}$, making $g(x) = \frac{1}{n}(\sum_{i=1}^{n-1} f_i(x) + f_{n1}(x) + f_{n2}(x))$ which can be written as $g(x) = \frac{1}{n+1}(\sum_{i=1}^{n-1} \frac{n+1}{n} f_i(x) + \frac{n+1}{n} f_{n1}(x) + \frac{n+1}{n} f_{n2}(x))$
    - $\frac{n+1}{n}|f_{n1}(x) - f_{n1}(y)| = \frac{n+1}{n}\frac{|f_n(x) - f_n(y)|}{2} \leq \frac{n+1}{2n} L_n ||x-y||^2$
    - Also increased probability of $f_n$ being sampled from $\frac{1}{n}$ to $\frac{2}{(n+1)}$
  - Duplicate each function $f_i$ a no. of times equal to it's Lipschitz constant of their gradient.
    - $g(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{L}{}_i \frac{f_i(x)}{L_i}$ which can be written as $g(x) = \frac{1}{\sum_k L_k}\sum_{i=1}^{n}\sum_{j=1}^{L}{}_i \frac{\sum_k L_k}{n}\frac{f_i(x)}{L_i}$. Now $L$ is equivalent to taking average of Lipschitz constants across $f_i'$ instead of taking max.

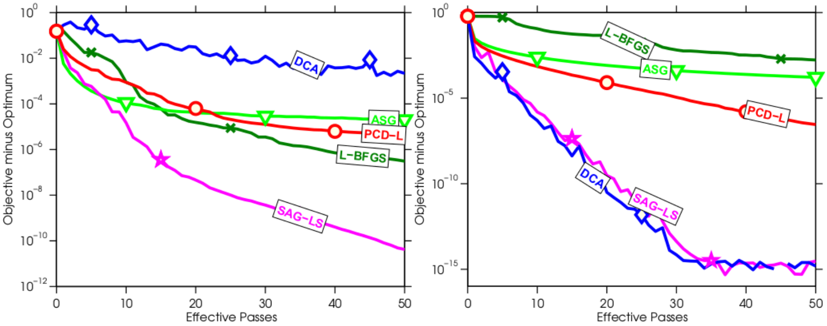
# Results-Comparison with FG & SG methods

- quantum ($n = 50000$, $p = 78$) and rcv1 ($n = 697641$, $p = 47236$)



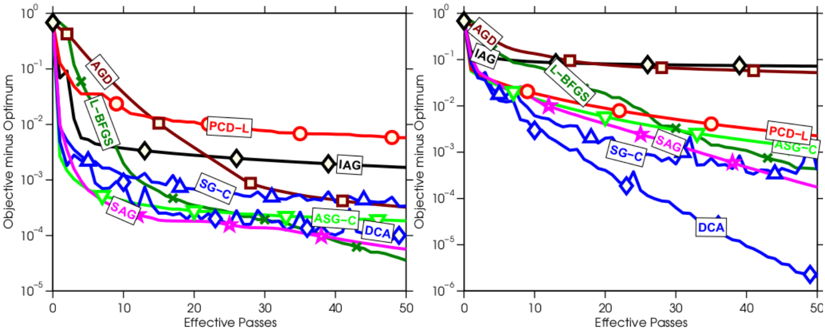
# Results-Comparison with CD methods

- quantum ($n = 50000$, $p = 78$) and rcv1 ($n = 697641$, $p = 47236$)



# Effect of non-uniform sampling

- protein ($n = 145751$, $p = 74$) and sido ($n = 12678$, $p = 4932$)

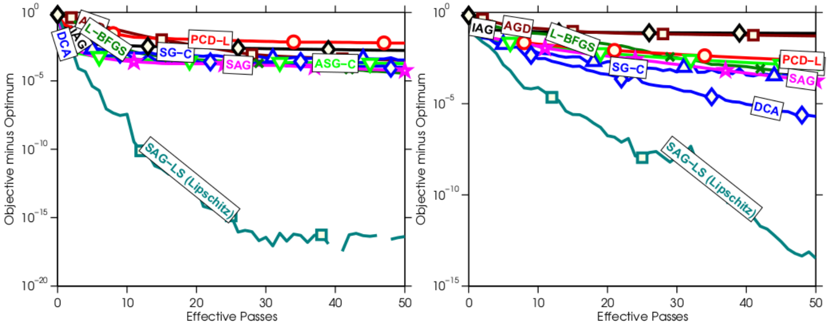



- Datasets where SAG had the worst relative performance.

Before non-uniform sampling.

# Effect of non-uniform sampling

- protein ($n = 145751$, $p = 74$) and sido ($n = 12678$, $p = 4932$)




- Lipschitz sampling helps a lot.

After non-uniform sampling.