# A Unified Theory of SGD: Variance Reduction, Sampling, Quantization and Coordinate Descent

The talk is based on the work [2] and slides [6]

Eduard Gorbunov[1] Filip Hanzely[2] Peter Richtárik[2 1]

[1]Moscow Institute of Physics and Technology, Russia
[2]King Abdullah University of Science and Technology, Saudi Arabia

INRIA, Paris, 18 October, 2019


# Table of Contents

# The Problem

$$\min_{x \in \mathbb{R}^d} f(x) + R(x) \tag{1}$$

# The Problem

$$\min_{x \in \mathbb{R}^d} f(x) + R(x) \tag{1}$$

- $f(x)$ — convex function with Lipschitz gradient.

# The Problem

$$\min_{x \in \mathbb{R}^d} f(x) + R(x) \tag{1}$$

- $f(x)$ — convex function with Lipschitz gradient.
- $R : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ — proximable (proper closed convex) regularizer

# Structure of $f$ in supervised learning theory and practice

We focus on the following situations

# Structure of $f$ in supervised learning theory and practice

We focus on the following situations

- $$f(x) = \mathbf{E}_{\xi \sim \mathcal{D}} \left[ f_\xi(x) \right] \tag{2}$$

# Structure of $f$ in supervised learning theory and practice

We focus on the following situations

- $$f(x) = \mathbf{E}_{\xi \sim \mathcal{D}} \left[ f_\xi(x) \right] \tag{2}$$

- $$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x). \tag{3}$$

# Structure of $f$ in supervised learning theory and practice

We focus on the following situations

- $$f(x) = \mathbf{E}_{\xi \sim \mathcal{D}} \left[ f_\xi(x) \right] \tag{2}$$
- $$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x). \tag{3}$$
- $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$ and

# Structure of $f$ in supervised learning theory and practice

We focus on the following situations

- $$f(x) = \mathbf{E}_{\xi \sim \mathcal{D}} \left[ f_\xi(x) \right] \tag{2}$$

- $$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x). \tag{3}$$

- $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$ and

$$f_i(x) = \frac{1}{m} \sum_{j=1}^{m} f_{ij}(x), \tag{4}$$

# Structure of $f$ in supervised learning theory and practice

We focus on the following situations

- $$f(x) = \mathbf{E}_{\xi \sim \mathcal{D}} \left[ f_\xi(x) \right] \tag{2}$$

- $$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x). \tag{3}$$

- $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$ and

$$f_i(x) = \frac{1}{m} \sum_{j=1}^{m} f_{ij}(x), \tag{4}$$

Typical case:

# Structure of $f$ in supervised learning theory and practice

We focus on the following situations

- $$f(x) = \mathbf{E}_{\xi \sim \mathcal{D}} \left[ f_\xi(x) \right] \tag{2}$$

- $$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x). \tag{3}$$

- $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$ and

$$f_i(x) = \frac{1}{m} \sum_{j=1}^{m} f_{ij}(x), \tag{4}$$

Typical case: Using the exact gradient of $\nabla f(x)$ is too expensive

# Structure of $f$ in supervised learning theory and practice

We focus on the following situations

- $$f(x) = \mathbf{E}_{\xi \sim \mathcal{D}} \left[ f_\xi(x) \right] \tag{2}$$

- $$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x). \tag{3}$$

- $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$ and

$$f_i(x) = \frac{1}{m} \sum_{j=1}^{m} f_{ij}(x), \tag{4}$$

Typical case: Using the exact gradient of $\nabla f(x)$ is too expensive, but an unbiased estimator of $\nabla f(x)$ can be computed efficiently.

# Popular approach to solve problem (1)

$$x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k). \tag{5}$$

# Popular approach to solve problem (1)

$$x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k). \tag{5}$$

- $g^k$ is unbiased gradient estimator: $\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k)$

# Popular approach to solve problem (1)

$$x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k). \tag{5}$$

- $g^k$ is unbiased gradient estimator: $\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k)$
- $\mathrm{prox}_R(x) \overset{\text{def}}{=} \mathrm{argmin}_u \left\{\frac{1}{2}\|u - x\|^2 + R(x)\right\}$

# Popular approach to solve problem (1)

$$x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k). \tag{5}$$

- $g^k$ is unbiased gradient estimator: $\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k)$
- $\mathrm{prox}_R(x) \overset{\text{def}}{=} \mathrm{argmin}_u \left\{\frac{1}{2}\|u - x\|^2 + R(x)\right\}$
- $\|\cdot\|$ — standard Euclidean norm

# Popular approach to solve problem (1)

$$x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k). \tag{5}$$

- $g^k$ is unbiased gradient estimator: $\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k)$
- $\mathrm{prox}_R(x) \overset{\text{def}}{=} \mathrm{argmin}_u \left\{\frac{1}{2}\|u - x\|^2 + R(x)\right\}$
- $\|\cdot\|$ — standard Euclidean norm

## The prox operator

- $x \to \mathrm{prox}_R(x)$ is a function

# Popular approach to solve problem (1)

$$x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k). \tag{5}$$

- $g^k$ is unbiased gradient estimator: $\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k)$
- $\mathrm{prox}_R(x) \overset{\text{def}}{=} \mathrm{argmin}_u \left\{\frac{1}{2}\|u - x\|^2 + R(x)\right\}$
- $\|\cdot\|$ — standard Euclidean norm

## The prox operator

- $x \to \mathrm{prox}_R(x)$ is a function
- $\|\mathrm{prox}_R(x) - \mathrm{prox}_R(y)\| \leq \|x - y\|$

# Stochastic gradient

There are infinitely many ways of getting unbiased estimator with "good" properties.

# Stochastic gradient

There are infinitely many ways of getting unbiased estimator with "good" properties.

- Flexibility to construct stochastic gradients in order to target desirable properties:

# Stochastic gradient

There are infinitely many ways of getting unbiased estimator with "good" properties.

- Flexibility to construct stochastic gradients in order to target desirable properties:
  - convergence speed

# Stochastic gradient

There are infinitely many ways of getting unbiased estimator with "good" properties.

- Flexibility to construct stochastic gradients in order to target desirable properties:
  - convergence speed
  - iteration cost

# Stochastic gradient

There are infinitely many ways of getting unbiased estimator with "good" properties.

- Flexibility to construct stochastic gradients in order to target desirable properties:
  - convergence speed
  - iteration cost
  - overall complexity

# Stochastic gradient

There are infinitely many ways of getting unbiased estimator with "good" properties.

- Flexibility to construct stochastic gradients in order to target desirable properties:
  - convergence speed
  - iteration cost
  - overall complexity
  - parallelizability

# Stochastic gradient

There are infinitely many ways of getting unbiased estimator with "good" properties.

- Flexibility to construct stochastic gradients in order to target desirable properties:
  - convergence speed
  - iteration cost
  - overall complexity
  - parallelizability
  - communication cost and etc.

# Stochastic gradient

There are infinitely many ways of getting unbiased estimator with "good" properties.

- Flexibility to construct stochastic gradients in order to target desirable properties:
  - convergence speed
  - iteration cost
  - overall complexity
  - parallelizability
  - communication cost and etc.
- Too many methods

# Stochastic gradient

There are infinitely many ways of getting unbiased estimator with "good" properties.

- Flexibility to construct stochastic gradients in order to target desirable properties:
  - convergence speed
  - iteration cost
  - overall complexity
  - parallelizability
  - communication cost and etc.
- Too many methods
  - Hard to keep up with new results

# Stochastic gradient

There are infinitely many ways of getting unbiased estimator with "good" properties.

- Flexibility to construct stochastic gradients in order to target desirable properties:
  - convergence speed
  - iteration cost
  - overall complexity
  - parallelizability
  - communication cost and etc.
- Too many methods
  - Hard to keep up with new results
  - Challenges in terms of analysis

# Stochastic gradient

There are infinitely many ways of getting unbiased estimator with "good" properties.

- Flexibility to construct stochastic gradients in order to target desirable properties:
  - convergence speed
  - iteration cost
  - overall complexity
  - parallelizability
  - communication cost and etc.
- Too many methods
  - Hard to keep up with new results
  - Challenges in terms of analysis
  - Some problems with fair comparison: different assumptions are used

# Bregman divergence

By $D_f(x, y)$ we denote the Bregman divergence associated with $f$:

$$D_f(x, y) \stackrel{\text{def}}{=} f(x) - f(y) - \langle \nabla f(y), x - y \rangle.$$

# Key assumption

## Assumption 1

Let $\{x^k\}$ be the random iterates produced by proximal SGD.

# Key assumption

## Assumption 1

Let $\{x^k\}$ be the random iterates produced by proximal SGD.

1. The stochastic gradients $g^k$ are unbiased

$$\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k) \quad \forall k \geq 0. \tag{6}$$

# Key assumption

## Assumption 1

Let $\{x^k\}$ be the random iterates produced by proximal SGD.

1. The stochastic gradients $g^k$ are unbiased

$$\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k) \quad \forall k \geq 0. \tag{6}$$

2. There exist non-negative constants $A, B, C, D_1, D_2, \rho$ and a (possibly) random sequence $\{\sigma_k^2\}_{k\geq 0}$ such that the following two relations hold

# Key assumption

## Assumption 1

Let $\{x^k\}$ be the random iterates produced by proximal SGD.

1. The stochastic gradients $g^k$ are unbiased

$$\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k) \quad \forall k \geq 0. \tag{6}$$

2. There exist non-negative constants $A, B, C, D_1, D_2, \rho$ and a (possibly) random sequence $\{\sigma_k^2\}_{k\geq 0}$ such that the following two relations hold

$$\mathbf{E}\left[\left\|g^k - \nabla f(x^*)\right\|^2 \mid x^k\right] \leq 2AD_f(x^k, x^*) + B\sigma_k^2 + D_1, \tag{7}$$

# Key assumption

## Assumption 1

Let $\{x^k\}$ be the random iterates produced by proximal SGD.

1. The stochastic gradients $g^k$ are unbiased

$$\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k) \quad \forall k \geq 0. \tag{6}$$

2. There exist non-negative constants $A, B, C, D_1, D_2, \rho$ and a (possibly) random sequence $\{\sigma_k^2\}_{k\geq 0}$ such that the following two relations hold

$$\mathbf{E}\left[\left\|g^k - \nabla f(x^*)\right\|^2 \mid x^k\right] \leq 2AD_f(x^k, x^*) + B\sigma_k^2 + D_1, \tag{7}$$

$$\mathbf{E}\left[\sigma_{k+1}^2 \mid \sigma_k^2\right] \leq (1-\rho)\sigma_k^2 + 2CD_f(x^k, x^*) + D_2, \quad \forall k \geq 0 \tag{8}$$

# Gradient Descent satisfies Assumption 1

## Assumption 2

Assume that $f$ is convex, i.e.

$$D_f(x, y) \geq 0 \quad \forall x, y \in \mathbb{R}^d,$$

and $L$-smooth, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^d.$$

# Gradient Descent satisfies Assumption 1

## Assumption 2

Assume that $f$ is convex, i.e.

$$D_f(x, y) \geq 0 \quad \forall x, y \in \mathbb{R}^d,$$

and $L$-smooth, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^d.$$

Assumption 2 implies that (see Nesterov's book [4])

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq 2LD_f(x, y) \quad \forall x, y \in \mathbb{R}^d.$$

# Gradient Descent satisfies Assumption 1

## Assumption 2

Assume that $f$ is convex, i.e.

$$D_f(x, y) \geq 0 \quad \forall x, y \in \mathbb{R}^d,$$

and $L$-smooth, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^d.$$

Assumption 2 implies that (see Nesterov's book [4])

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq 2LD_f(x, y) \quad \forall x, y \in \mathbb{R}^d.$$

Therefore, if $f$ satisfies Assumption 2, then gradient descent satisfies Assumption 1 with

$$A = L, B = 0, D_1 = 0, \sigma_k = 0, \rho = 1, C = 0, D_2 = 0.$$

# Other assumptions

## Assumption 3 (Unique solution)

The problem (1) has an unique minimizer $x^*$.

# Other assumptions

## Assumption 3 (Unique solution)

The problem (1) has an unique minimizer $x^*$.

## Assumption 4 ($\mu$-strong quasi-convexity)

There exists $\mu > 0$ such that $f : \mathbb{R}^d \to \mathbb{R}$ is $\mu$-strongly quasi-convex, i.e.

$$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{\mu}{2} \|x^* - x\|^2, \qquad \forall x \in \mathbb{R}^d. \tag{9}$$

# Main result

## Theorem 1

Let Assumptions 1, 3 and 4 be satisfied.

# Main result

## Theorem 1

Let Assumptions 1, 3 and 4 be satisfied. Choose constant $M$ such that $M > \frac{B}{\rho}$. Choose a stepsize satisfying

$$0 < \gamma \leq \min\left\{\frac{1}{\mu}, \frac{1}{A + CM}\right\}.$$

# Main result

## Theorem 1

Let Assumptions 1, 3 and 4 be satisfied. Choose constant $M$ such that $M > \frac{B}{\rho}$. Choose a stepsize satisfying

$$0 < \gamma \leq \min\left\{\frac{1}{\mu}, \frac{1}{A + CM}\right\}. \tag{10}$$

Then the iterates $\{x^k\}_{k\geq 0}$ of proximal SGD satisfy

$$\mathbf{E}\left[V^k\right] \leq \max\left\{(1-\gamma\mu)^k, \left(1 + \frac{B}{M} - \rho\right)^k\right\} V^0 + \frac{(D_1 + MD_2)\gamma^2}{\min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}}, \tag{11}$$

# Main result

## Theorem 1

Let Assumptions 1, 3 and 4 be satisfied. Choose constant $M$ such that $M > \frac{B}{\rho}$. Choose a stepsize satisfying

$$0 < \gamma \leq \min\left\{\frac{1}{\mu}, \frac{1}{A + CM}\right\}. \tag{10}$$

Then the iterates $\{x^k\}_{k\geq 0}$ of proximal SGD satisfy

$$\mathbf{E}\left[V^k\right] \leq \max\left\{(1-\gamma\mu)^k, \left(1 + \frac{B}{M} - \rho\right)^k\right\} V^0 + \frac{(D_1 + MD_2)\gamma^2}{\min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}}, \tag{11}$$

where the Lyapunov function $V^k$ is defined by $V^k \stackrel{\text{def}}{=} \left\|x^k - x^*\right\|^2 + M\gamma^2\sigma_k^2$.

# Methods

| Problem | Method | Alg # | Citation | VR? | AS? | Quant? | RCD? | Section | Result |
|---|---|---|---|---|---|---|---|---|---|
| (1)+(2) | SGD | Alg 1 | [26] | ✗ | ✗ | ✗ | ✗ | A.1 | Cor A.1 |
| (1)+(3) | SGD-SR | Alg 2 | [6] | ✗ | ✓ | ✗ | ✗ | A.2 | Cor A.2 |
| (1)+(3) | SGD-MB | Alg 3 | **NEW** | ✗ | ✗ | ✗ | ✗ | A.3 | Cor A.3 |
| (1)+(3) | SGD-star | Alg 4 | **NEW** | ✓ | ✓ | ✗ | ✗ | A.4 | Cor A.4 |
| (1)+(3) | SAGA | Alg 5 | [5] | ✓ | ✗ | ✗ | ✗ | A.5 | Cor A.5 |
| (1)+(3) | N-SAGA | Alg 6 | **NEW** | ✗ | ✗ | ✗ | ✗ | A.6 | Cor A.6 |
| (1) | SEGA | Alg 7 | [11] | ✓ | ✗ | ✗ | ✓ | A.7 | Cor A.7 |
| (1) | N-SEGA | Alg 8 | **NEW** | ✗ | ✗ | ✗ | ✓ | A.8 | Cor A.8 |
| (1)+(3) | SVRG[a] | Alg 9 | [15] | ✓ | ✗ | ✗ | ✗ | A.9 | Cor A.9 |
| (1)+(3) | L-SVRG | Alg 10 | [13, 18] | ✓ | ✗ | ✗ | ✗ | A.10 | Cor A.10 |
| (1)+(3) | DIANA | Alg 11 | [20, 14] | ✗ | ✗ | ✓ | ✗ | A.11 | Cor A.11 |
| (1)+(3) | DIANA[b] | Alg 12 | [20, 14] | ✓ | ✗ | ✓ | ✗ | A.11 | Cor A.12 |
| (1)+(3) | Q-SGD-SR | Alg 13 | **NEW** | ✗ | ✓ | ✓ | ✗ | A.12 | Cor A.13 |
| (1)+(3)+(4) | VR-DIANA | Alg 14 | [14] | ✓ | ✗ | ✓ | ✗ | A.13 | Cor A.15 |
| (1)+(3) | JacSketch | Alg 15 | [9] | ✓ | ✓✗ | ✗ | ✗ | A.14 | Cor A.16 |

Table 1: List of specific existing (in some cases generalized) and new methods which fit our general analysis framework. VR = variance reduced method, AS = arbitrary sampling, Quant = supports gradient quantization, RCD = randomized coordinate descent type method. [a] Special case of SVRG with 1 outer loop only; [b] Special case of DIANA with 1 node and quantization of exact gradient.

# Parameters

| Method | $A$ | $B$ | $\rho$ | $C$ | $D_1$ | $D_2$ |
|---|---|---|---|---|---|---|
| SGD | $2L$ | $0$ | $1$ | $0$ | $2\sigma^2$ | $0$ |
| SGD-SR | $2\mathcal{L}$ | $0$ | $1$ | $0$ | $2\sigma^2$ | $0$ |
| SGD-MB | $\frac{A'+L(\tau-1)}{\tau}$ | $0$ | $1$ | $0$ | $\frac{D'}{\tau}$ | $0$ |
| SGD-star | $2\mathcal{L}$ | $0$ | $1$ | $0$ | $0$ | $0$ |
| SAGA | $2L$ | $2$ | $1/n$ | $L/n$ | $0$ | $0$ |
| N-SAGA | $2L$ | $2$ | $1/n$ | $L/n$ | $2\sigma^2$ | $\frac{\sigma^2}{n}$ |
| SEGA | $2dL$ | $2d$ | $1/d$ | $L/d$ | $0$ | $0$ |
| N-SEGA | $2dL$ | $2d$ | $1/d$ | $L/d$ | $2d\sigma^2$ | $\frac{\sigma^2}{d}$ |
| SVRG[a] | $2L$ | $2$ | $0$ | $0$ | $0$ | $0$ |
| L-SVRG | $2L$ | $2$ | $p$ | $Lp$ | $0$ | $0$ |
| DIANA | $\left(1+\frac{2\omega}{n}\right)L$ | $\frac{2\omega}{n}$ | $\alpha$ | $L\alpha$ | $\frac{(1+\omega)\sigma^2}{n}$ | $\alpha\sigma^2$ |
| DIANA[b] | $(1+2\omega)L$ | $2\omega$ | $\alpha$ | $L\alpha$ | $0$ | $0$ |
| Q-SGD-SR | $2(1+\omega)\mathcal{L}$ | $0$ | $1$ | $0$ | $2(1+\omega)\sigma^2$ | $0$ |
| VR-DIANA | $\left(1+\frac{4\omega+2}{n}\right)L$ | $\frac{2(\omega+1)}{n}$ | $\alpha$ | $\left(\frac{1}{m}+4\alpha\right)L$ | $0$ | $0$ |
| JacSketch | $2\mathcal{L}_1$ | $\frac{2\lambda_{\max}}{n}$ | $\lambda_{\min}$ | $\frac{\mathcal{L}_2}{n}$ | $0$ | $0$ |

# Proximal Gradient Descent

We checked that GD satisfies Assumption 1 when $f$ is convex and $L$-smooth with

$$A = L, B = 0, D_1 = 0, \sigma_k = 0, \rho = 1, C = 0, D_2 = 0.$$

# Proximal Gradient Descent

We checked that GD satisfies Assumption 1 when $f$ is convex and $L$-smooth with

$$A = L, B = 0, D_1 = 0, \sigma_k = 0, \rho = 1, C = 0, D_2 = 0.$$

We can choose $M = 1$ in Theorem 1 and get

# Proximal Gradient Descent

We checked that GD satisfies Assumption 1 when $f$ is convex and $L$-smooth with

$$A = L, B = 0, D_1 = 0, \sigma_k = 0, \rho = 1, C = 0, D_2 = 0.$$

We can choose $M = 1$ in Theorem 1 and get

- $0 < \gamma \leq \frac{1}{L}$ (since $\mu \leq L$ and $C = 0$)

# Proximal Gradient Descent

We checked that GD satisfies Assumption 1 when $f$ is convex and $L$-smooth with

$$A = L, B = 0, D_1 = 0, \sigma_k = 0, \rho = 1, C = 0, D_2 = 0.$$

We can choose $M = 1$ in Theorem 1 and get

- $0 < \gamma \leq \frac{1}{L}$ (since $\mu \leq L$ and $C = 0$)
- $V^k = \|x^k - x^*\|^2$ (since $\sigma_k = 0$)

# Proximal Gradient Descent

We checked that GD satisfies Assumption 1 when $f$ is convex and $L$-smooth with

$$A = L, B = 0, D_1 = 0, \sigma_k = 0, \rho = 1, C = 0, D_2 = 0.$$

We can choose $M = 1$ in Theorem 1 and get

- $0 < \gamma \leq \frac{1}{L}$ (since $\mu \leq L$ and $C = 0$)
- $V^k = \|x^k - x^*\|^2$ (since $\sigma_k = 0$)
- The rate: $\|x^k - x^*\|^2 \leq (1 - \gamma\mu)^k \|x^0 - x^*\|^2$

# Proximal Gradient Descent

We checked that GD satisfies Assumption 1 when $f$ is convex and $L$-smooth with

$$A = L, B = 0, D_1 = 0, \sigma_k = 0, \rho = 1, C = 0, D_2 = 0.$$

We can choose $M = 1$ in Theorem 1 and get

- $0 < \gamma \leq \frac{1}{L}$ (since $\mu \leq L$ and $C = 0$)
- $V^k = \|x^k - x^*\|^2$ (since $\sigma_k = 0$)
- The rate: $\|x^k - x^*\|^2 \leq (1 - \gamma\mu)^k \|x^0 - x^*\|^2$
- In particular, for $\gamma = \frac{1}{L}$ we recover the standard rate for proximal gradient descent:

$$k \geq \frac{L}{\mu} \log \frac{1}{\varepsilon} \quad \Longrightarrow \quad \|x^k - x^*\|^2 \leq \varepsilon \|x^0 - x^*\|^2.$$

# SGD-SR (see Gower et al. (2019), [3])

## Stochastic reformulation

$$\min_{x\in\mathbb{R}^d} f(x) + R(x), \quad f(x) = \mathbf{E}_{\mathcal{D}}\left[f_\xi(x)\right], \quad f_\xi(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \xi_i f_i(x) \tag{12}$$

# SGD-SR (see Gower et al. (2019), [3])

## Stochastic reformulation

$$\min_{x\in\mathbb{R}^d} f(x) + R(x), \quad f(x) = \mathbf{E}_{\mathcal{D}}\left[f_\xi(x)\right], \quad f_\xi(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \xi_i f_i(x) \tag{12}$$

1. $\xi \sim \mathcal{D}$: $\mathbf{E}_{\mathcal{D}}\left[\xi_i\right] = 1$ for all $i$

# SGD-SR (see Gower et al. (2019), [3])

## Stochastic reformulation

$$\min_{x\in\mathbb{R}^d} f(x)+R(x), \quad f(x)=\mathbf{E}_{\mathcal{D}}\left[f_\xi(x)\right], \quad f_\xi(x)\overset{\text{def}}{=}\frac{1}{n}\sum_{i=1}^{n}\xi_i f_i(x) \tag{12}$$

1. $\xi\sim\mathcal{D}$: $\mathbf{E}_{\mathcal{D}}\left[\xi_i\right]=1$ for all $i$
2. $f_i$ (for all $i$) is smooth, possibly non-convex function

# SGD-SR (see Gower et al. (2019), [3])

## Stochastic reformulation

$$\min_{x\in\mathbb{R}^d} f(x)+R(x), \quad f(x)=\mathbf{E}_{\mathcal{D}}\left[f_\xi(x)\right], \quad f_\xi(x)\overset{\text{def}}{=}\frac{1}{n}\sum_{i=1}^{n}\xi_i f_i(x) \tag{12}$$

1. $\xi\sim\mathcal{D}$: $\mathbf{E}_{\mathcal{D}}\left[\xi_i\right]=1$ for all $i$
2. $f_i$ (for all $i$) is smooth, possibly non-convex function

**Algorithm 4** SGD-SR

**Input:** learning rate $\gamma>0$, starting point $x^0\in\mathbb{R}^d$, distribution $\mathcal{D}$ over $\xi\in\mathbb{R}^n$ such that $\mathbf{E}_{\mathcal{D}}\left[\xi\right]$ is vector of ones

1: **for** $k=0,1,2,\ldots$ **do**
2: Sample $\xi\sim\mathcal{D}$
3: $g^k=\nabla f_\xi(x^k)$
4: $x^{k+1}=\mathrm{prox}_{\gamma R}(x^k-\gamma g^k)$
5: **end for**

# SGD-SR (see Gower et al. (2019), [3])

## Expected smoothness

We say that $f$ is $\mathcal{L}$-smooth in expectation with respect to distribution $\mathcal{D}$ if there exists $\mathcal{L} = \mathcal{L}(f, \mathcal{D}) > 0$ such that

$$\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(x) - \nabla f_\xi(x^*)\|^2\right] \leq 2\mathcal{L}D_f(x, x^*), \tag{13}$$

for all $x \in \mathbb{R}^d$ and write $(f, \mathcal{D}) \sim ES(\mathcal{L})$.

# SGD-SR (see Gower et al. (2019), [3])

## Lemma 1 (Generalization of Lemma 2.4, [3])

If $(f, \mathcal{D}) \sim ES(\mathcal{L})$, then

$$\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(x) - \nabla f(x^*)\|^2\right] \leq 4\mathcal{L}D_f(x, x^*) + 2\sigma^2. \tag{14}$$

where $\sigma^2 \overset{\text{def}}{=} \mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(x^*) - \nabla f(x^*)\|^2\right]$.

# SGD-SR (see Gower et al. (2019), [3])

## Lemma 1 (Generalization of Lemma 2.4, [3])

If $(f, \mathcal{D}) \sim ES(\mathcal{L})$, then

$$\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(x) - \nabla f(x^*)\|^2\right] \leq 4\mathcal{L}D_f(x, x^*) + 2\sigma^2. \tag{14}$$

where $\sigma^2 \overset{\text{def}}{=} \mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(x^*) - \nabla f(x^*)\|^2\right]$.

That is, SGD-SR satisfies Assumption 1 with

$$A = 2\mathcal{L}, B = 0, D_1 = 2\sigma^2, \sigma_k = 0, \rho = 1, C = 0, D_2 = 0.$$

# SGD-SR (see Gower et al. (2019), [3])

## Lemma 1 (Generalization of Lemma 2.4, [3])

If $(f, \mathcal{D}) \sim ES(\mathcal{L})$, then

$$\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(x) - \nabla f(x^*)\|^2\right] \leq 4\mathcal{L}D_f(x, x^*) + 2\sigma^2. \tag{14}$$

where $\sigma^2 \stackrel{\text{def}}{=} \mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(x^*) - \nabla f(x^*)\|^2\right]$.

That is, SGD-SR satisfies Assumption 1 with

$$A = 2\mathcal{L}, B = 0, D_1 = 2\sigma^2, \sigma_k = 0, \rho = 1, C = 0, D_2 = 0.$$

If $\gamma^k \equiv \gamma \leq \frac{1}{2\mathcal{L}}$ then Theorem 1 implies that

$$\mathbf{E}\|x^k - x^*\|^2 \leq (1 - \gamma\mu)^k\|x^0 - x^*\|^2 + \frac{2\gamma\sigma^2}{\mu}.$$

# SAGA (see Defazio, Bach & Lacoste-Julien (2014), [1])

Consider the finite-sum minimization problem

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) + R(x), \tag{15}$$

# SAGA (see Defazio, Bach & Lacoste-Julien (2014), [1])

Consider the finite-sum minimization problem

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) + R(x), \tag{15}$$

where $f_i$ is convex, $L$-smooth for each $i$ and $f$ is $\mu$-strongly convex.

# SAGA (see Defazio, Bach & Lacoste-Julien (2014), [1])

Consider the finite-sum minimization problem

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) + R(x), \tag{15}$$

where $f_i$ is convex, $L$-smooth for each $i$ and $f$ is $\mu$-strongly convex.

**Algorithm 7** SAGA [1]

**Input:** learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$

1: Set $\psi_j^0 = x^0$ for each $j \in [n]$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Sample $j \in [n]$ uniformly at random
4: Set $\phi_j^{k+1} = x^k$ and $\phi_i^{k+1} = \phi_i^k$ for $i \neq j$
5: $g^k = \nabla f_j(\phi_j^{k+1}) - \nabla f_j(\phi_j^k) + \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\phi_i^k)$
6: $x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$
7: **end for**

# SAGA (see Defazio, Bach & Lacoste-Julien (2014), [1])

## Lemma 2

We have

$$\mathbf{E}\left[\left\|g^k - \nabla f(x^*)\right\|^2 \mid x^k\right] \leq 4LD_f(x^k, x^*) + 2\sigma_k^2$$

# SAGA (see Defazio, Bach & Lacoste-Julien (2014), [1])

## Lemma 2

We have

$$\mathbf{E}\left[\left\|g^k - \nabla f(x^*)\right\|^2 \mid x^k\right] \leq 4LD_f(x^k, x^*) + 2\sigma_k^2 \tag{16}$$

and

$$\mathbf{E}\left[\sigma_{k+1}^2 \mid x^k\right] \leq \left(1 - \frac{1}{n}\right)\sigma_k^2 + \frac{2L}{n}D_f(x^k, x^*), \tag{17}$$

# SAGA (see Defazio, Bach & Lacoste-Julien (2014), [1])

## Lemma 2

We have

$$\mathbf{E}\left[\left\|g^k - \nabla f(x^*)\right\|^2 \mid x^k\right] \leq 4LD_f(x^k, x^*) + 2\sigma_k^2 \tag{16}$$

and

$$\mathbf{E}\left[\sigma_{k+1}^2 \mid x^k\right] \leq \left(1 - \frac{1}{n}\right)\sigma_k^2 + \frac{2L}{n}D_f(x^k, x^*), \tag{17}$$

where $\sigma_k^2 = \frac{1}{n}\sum\limits_{i=1}^{n}\left\|\nabla f_i(\phi_i^k) - \nabla f_i(x^*)\right\|^2$.

# SAGA (see Defazio, Bach & Lacoste-Julien (2014), [1])

That is, SAGA satisfies Assumption 1 with

$$A = 2L, B = 1, D_1 = 0, \rho = \frac{1}{n}, C = \frac{L}{n}, D_2 = 0.$$

# SAGA (see Defazio, Bach & Lacoste-Julien (2014), [1])

That is, SAGA satisfies Assumption 1 with

$$A = 2L, B = 1, D_1 = 0, \rho = \frac{1}{n}, C = \frac{L}{n}, D_2 = 0.$$

Theorem 1 with $M = 4n$ implies that for $\gamma = \frac{1}{6L}$

$$\mathbf{E}V^k \leq \left(1 - \min\left\{\frac{\mu}{6L}, \frac{1}{2n}\right\}\right)^2 V^0.$$

# SGD-star

Consider the same situation as for SGD-SR. Recall that for SGD-SR we got

$$\mathbf{E}\|x^k - x^*\|^2 \leq (1-\gamma\mu)^k\|x^0 - x^*\|^2 + \frac{2\gamma\sigma^2}{\mu}, \quad \gamma \in \left(0, \frac{1}{2\mathcal{L}}\right]$$

# SGD-star

Consider the same situation as for SGD-SR. Recall that for SGD-SR we got

$$\mathbf{E}\|x^k - x^*\|^2 \leq (1-\gamma\mu)^k\|x^0 - x^*\|^2 + \frac{2\gamma\sigma^2}{\mu}, \quad \gamma \in \left(0, \frac{1}{2\mathcal{L}}\right]$$

**Algorithm 9** SGD-star

**Input:** learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, distribution $\mathcal{D}$ over $\xi \in \mathbb{R}^n$ such that $\mathbf{E}_{\mathcal{D}}[\xi]$ is vector of ones

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: Sample $\xi \sim \mathcal{D}$
3: $g^k = \nabla f_\xi(x^k) - \nabla f_\xi(x^*) + \nabla f(x^*)$
4: $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$
5: **end for**

# SGD-star

## Lemma 3 (Lemma 2.4 from [3])

If $(f, \mathcal{D}) \sim ES(\mathcal{L})$, then

$$\mathbf{E}_{\mathcal{D}}\left[\left\|g^k - \nabla f(x^*)\right\|^2\right] \leq 4\mathcal{L}D_f(x^k, x^*). \tag{18}$$

Thus, SGD-star satisfies Assumption 1 with

$$A = 2\mathcal{L}, B = 0, D_1 = 0, \sigma_k = 0, \rho = 1, C = 0, D_2 = 0.$$

# SGD-star

## Lemma 3 (Lemma 2.4 from [3])

If $(f, \mathcal{D}) \sim ES(\mathcal{L})$, then

$$\mathbf{E}_{\mathcal{D}}\left[\left\|g^k - \nabla f(x^*)\right\|^2\right] \leq 4\mathcal{L}D_f(x^k, x^*). \tag{18}$$

Thus, SGD-star satisfies Assumption 1 with

$$A = 2\mathcal{L}, B = 0, D_1 = 0, \sigma_k = 0, \rho = 1, C = 0, D_2 = 0.$$

If $\gamma^k \equiv \gamma \leq \frac{1}{2\mathcal{L}}$ then Theorem 1 implies that

$$\mathbf{E}\|x^k - x^*\|^2 \leq (1 - \gamma\mu)^k \|x^0 - x^*\|^2.$$

# Quantitative definition of variance reduction

## Variance-reduced methods

We say that SGD with stochastic gradients satisfying Assumption 1 is variance-reduced if $D_1 = D_2 = 0$.

# Quantitative definition of variance reduction

## Variance-reduced methods

We say that SGD with stochastic gradients satisfying Assumption 1 is variance-reduced if $D_1 = D_2 = 0$.

- Variance-reduced methods converge to the exact solution

# Quantitative definition of variance reduction

## Variance-reduced methods

We say that SGD with stochastic gradients satisfying Assumption 1 is variance-reduced if $D_1 = D_2 = 0$.

- Variance-reduced methods converge to the exact solution
- The sequence $\{\sigma_k^2\}_{k\geq 0}$ reflects the progress of the variance reduction process.

# Limitations and Extensions

1. We consider only $L$-smooth $\mu$-strongly quasi-convex case.

# Limitations and Extensions

1. We consider only $L$-smooth $\mu$-strongly quasi-convex case.
2. It would interesting to unify the theory for biased gradients estimator (e.g. `SAG` [7], `SARAH` [5], zero-order optimization and etc.)

# Limitations and Extensions

1. We consider only $L$-smooth $\mu$-strongly quasi-convex case.
2. It would interesting to unify the theory for biased gradients estimator (e.g. SAG [7], SARAH [5], zero-order optimization and etc.)
3. Our analysis doesn't recover the best known rates for RCD type methods with importance sampling.

# Limitations and Extensions

1. We consider only $L$-smooth $\mu$-strongly quasi-convex case.
2. It would interesting to unify the theory for biased gradients estimator (e.g. SAG [7], SARAH [5], zero-order optimization and etc.)
3. Our analysis doesn't recover the best known rates for RCD type methods with importance sampling.
4. An extension of iteration dependent parameters $A, B, C, D_1, D_2, \rho$ would cover a new methods, such as SGD with decreasing stepsizes.

# Limitations and Extensions

1. We consider only $L$-smooth $\mu$-strongly quasi-convex case.
2. It would interesting to unify the theory for biased gradients estimator (e.g. SAG [7], SARAH [5], zero-order optimization and etc.)
3. Our analysis doesn't recover the best known rates for RCD type methods with importance sampling.
4. An extension of iteration dependent parameters $A, B, C, D_1, D_2, \rho$ would cover a new methods, such as SGD with decreasing stepsizes.
5. We consider only non-accelerated methods. It would be interesting to provide an unified analysis of stochastic methods with acceleration and momentum.

# Bibliography I


Aaron Defazio, Francis Bach, and Simon Lacoste-Julien.
SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives.
In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.


Eduard Gorbunov, Filip Hanzely, and Peter Richtárik.
A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent.
*arXiv preprint arXiv:1905.11261*, 2019.

# Bibliography II



Robert M Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik.
SGD: General analysis and improved rates.
*arXiv preprint arXiv:1901.09401*, 2019.


Yurii Nesterov.
Introductory lectures on convex optimization: a basic course.
2004.


Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč.
SARAH: A novel method for machine learning problems using stochastic recursive gradient.
In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2613–2621. PMLR, 2017.

# Bibliography III


Peter Richtárik.
A guided walk through the zoo of stochastic gradient descent methods.

*ICCOPT 2019 Summer School*.


Nicolas Le Roux, Mark Schmidt, and Francis Bach.
A stochastic gradient method with an exponential convergence rate for finite training sets.
In *Advances in Neural Information Processing Systems*, pages 2663–2671, 2012.