# PAGE: A Simple and Optimal Probabilistic Gradient Estimator for Nonconvex Optimization

**Zhize Li**

King Abdullah University of Science and Technology (KAUST)

https://zhizeli.github.io

ICML 2021

King Abdullah University of Science and Technology

ICML | 2021

Thirty-eighth International Conference on Machine Learning

# Joint work with

Hongyan Bao

Xiangliang Zhang

Peter Richtárik

# Overview

# Problem

We consider the general nonconvex optimization problem

$$\min_{x\in\mathbb{R}^d} f(x),$$

where the nonconvex function $f$ has the following two forms:

- **Finite-sum form**:

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

  ($n$ data samples, $f_i$ is the nonconvex loss on data $i$)

- **Online form**:

$$f(x) := \mathbb{E}_{\zeta\sim\mathcal{D}}[F(x,\zeta)].$$

  (data is drawn from an unknown distribution $\mathcal{D}$)

# Related Work

There exist many methods for solving this optimization problem with both forms, such as Gradient Descent (GD), Stochastic GD (SGD), and many variance-reduced methods (e.g., SVRG, SVRG+, L-SVRG, SAGA, SCSG, SNVRG, SARAH, SPIDER, SpiderBoost and SSRGD).

However, these methods either *do not achieve the optimal results* or are *complicated in algorithmic structure and/or convergence analysis*.

# Related Work

There exist many methods for solving this optimization problem with both forms, such as Gradient Descent (GD), Stochastic GD (SGD), and many variance-reduced methods (e.g., SVRG, SVRG+, L-SVRG, SAGA, SCSG, SNVRG, SARAH, SPIDER, SpiderBoost and SSRGD).

However, these methods either *do not achieve the optimal results* or are *complicated in algorithmic structure and/or convergence analysis*.

In this work, we provide a simple PAGE algorithm for achieving optimal results with simple convergence analysis, and provide tight lower bounds for validating the optimality.

# Our PAGE Algorithm

**Algorithm 1** ProbAbilistic Gradient Estimator (PAGE)

**Input:** initial $x^0$, stepsize $\eta$, minibatch $b$, $b'$, probability $\{p_t\}$

1: $g^0 = \frac{1}{b}\sum_{i\in I}\nabla f_i(x^0)$ // $I$ denotes random minibatch samples with $|I| = b$

2: **for** $t = 0, 1, 2, \ldots$ **do**

3: $x^{t+1} = x^t - \eta g^t$

4: $g^{t+1} = \begin{cases} \frac{1}{\boldsymbol{b}}\sum_{i\in I}\nabla f_i(x^{t+1}) & \text{with prob. } p_t \\ g^t + \frac{1}{\boldsymbol{b'}}\sum_{i\in I'}(\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) & \text{with prob. } 1 - p_t \end{cases}$

5: **end for**

**Output:** $\hat{x}_T$ chosen uniformly from $\{x^t\}_{t\in[T]}$

- PAGE uses minibatch SGD update with probability $p_t$, and reuses the previous gradient $g^t$ with a small adjustment (lower computational cost if $\boldsymbol{b'} \ll \boldsymbol{b}$) with probability $1 - p_t$.

# Convergence Results (finite-sum)

Average $L$-smooth: $\mathbb{E}_i[\|\nabla f_i(x) - \nabla f_i(y)\|^2] \le L^2\|x - y\|^2$

**Theorem 1 (Optimal result of PAGE in finite-sum case)**

*Suppose $f$ is average $L$-smooth, choosing appropriate parameters, the number of stochastic gradient computations of PAGE for finding an $\epsilon$-approximate solution $\mathbb{E}[\|\nabla f(\hat{x}_T)\|] \le \epsilon$ is $\#\mathrm{grad} = \boldsymbol{O(n + \frac{\sqrt{n}L}{\epsilon^2})}$.*

# Proof Sketch of Theorem 1

**Lemma 1** (one iteration). Suppose $f$ is $L$-smooth and $x^{t+1} := x^t - \eta g^t$. Then for any $g^t \in \mathbb{R}^d$ and $\eta > 0$, we have

$$f(x^{t+1}) \leq f(x^t) - \frac{\eta}{2}\|\nabla f(x^t)\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|x^{t+1} - x^t\|^2 + \frac{\eta}{2}\|g^t - \nabla f(x^t)\|^2. \quad (1)$$

# Proof Sketch of Theorem 1

**Lemma 1** (one iteration). Suppose $f$ is $L$-smooth and $x^{t+1} := x^t - \eta g^t$. Then for any $g^t \in \mathbb{R}^d$ and $\eta > 0$, we have

$$f(x^{t+1}) \le f(x^t) - \frac{\eta}{2}\|\nabla f(x^t)\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|x^{t+1} - x^t\|^2 + \frac{\eta}{2}\|g^t - \nabla f(x^t)\|^2. \quad (1)$$

**Lemma 2** (variance). Under average $L$-smoothness assumption, the gradient estimator $g^{t+1}$ of PAGE (Line 4 of Algorithm 1) satisfies

$$\mathbb{E}\left[\|g^{t+1} - \nabla f(x^{t+1})\|^2\right] \le (1 - p_t)\|g^t - \nabla f(x^t)\|^2 + \frac{(1 - p_t)L^2}{b'}\|x^{t+1} - x^t\|^2. \quad (2)$$

# Proof Sketch of Theorem 1

**Lemma 1** (one iteration). Suppose $f$ is $L$-smooth and $x^{t+1} := x^t - \eta g^t$. Then for any $g^t \in \mathbb{R}^d$ and $\eta > 0$, we have

$$f(x^{t+1}) \le f(x^t) - \frac{\eta}{2}\|\nabla f(x^t)\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|x^{t+1} - x^t\|^2 + \frac{\eta}{2}\|g^t - \nabla f(x^t)\|^2. \quad (1)$$

**Lemma 2** (variance). Under average $L$-smoothness assumption, the gradient estimator $g^{t+1}$ of PAGE (Line 4 of Algorithm 1) satisfies

$$\mathbb{E}\left[\|g^{t+1} - \nabla f(x^{t+1})\|^2\right] \le (1-p_t)\|g^t - \nabla f(x^t)\|^2 + \frac{(1-p_t)L^2}{b'}\|x^{t+1} - x^t\|^2. \quad (2)$$

Adding (1) with $\frac{\eta}{2p} \times$ (2) and letting $\Phi_t := f(x^t) - f^* + \frac{\eta}{2p}\|g^t - \nabla f(x^t)\|^2$, we have $\mathbb{E}[\Phi_{t+1}] \le \mathbb{E}\left[\Phi_t - \frac{\eta}{2}\|\nabla f(x^t)\|^2\right]$. Summing up from $t = 0$ to $T-1$, we get $\mathbb{E}[\Phi_T] \le \mathbb{E}[\Phi_0] - \frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(x^t)\|^2]$.

# Proof Sketch of Theorem 1

**Lemma 1** (one iteration). Suppose $f$ is $L$-smooth and $x^{t+1} := x^t - \eta g^t$. Then for any $g^t \in \mathbb{R}^d$ and $\eta > 0$, we have

$$f(x^{t+1}) \le f(x^t) - \frac{\eta}{2}\|\nabla f(x^t)\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|x^{t+1} - x^t\|^2 + \frac{\eta}{2}\|g^t - \nabla f(x^t)\|^2. \quad (1)$$

**Lemma 2** (variance). Under average $L$-smoothness assumption, the gradient estimator $g^{t+1}$ of PAGE (Line 4 of Algorithm 1) satisfies

$$\mathbb{E}\left[\|g^{t+1} - \nabla f(x^{t+1})\|^2\right] \le (1 - p_t)\|g^t - \nabla f(x^t)\|^2 + \frac{(1-p_t)L^2}{b'}\|x^{t+1} - x^t\|^2. \quad (2)$$

Adding (1) with $\frac{\eta}{2p} \times$ (2) and letting $\Phi_t := f(x^t) - f^* + \frac{\eta}{2p}\|g^t - \nabla f(x^t)\|^2$, we have $\mathbb{E}\left[\Phi_{t+1}\right] \le \mathbb{E}\left[\Phi_t - \frac{\eta}{2}\|\nabla f(x^t)\|^2\right]$. Summing up from $t = 0$ to $T - 1$, we get $\mathbb{E}[\Phi_T] \le \mathbb{E}[\Phi_0] - \frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(x^t)\|^2]$.

Then according to the random output $\hat{x}_T$ of PAGE, we have

$$\mathbb{E}[\|\nabla f(\hat{x}_T)\|^2] \le \frac{2\Phi_0}{\eta T} = \epsilon^2, \quad T = \frac{2\Phi_0}{\epsilon^2\eta}, \quad \#\text{grad} = b + T(pb + (1-p)b') = O\left(n + \frac{\sqrt{n}L}{\epsilon^2}\right).$$

# Convergence Result and Lower Bound (finite-sum)

Average $L$-smooth: $\mathbb{E}_i[\|\nabla f_i(x) - \nabla f_i(y)\|^2] \le L^2\|x-y\|^2$

**Theorem 1 (Optimal result of PAGE in finite-sum case)**

*Suppose $f$ is average $L$-smooth, choosing appropriate parameters, the number of stochastic gradient computations of PAGE for finding an $\epsilon$-approximate solution $\mathbb{E}[\|\nabla f(\hat{x}_T)\|] \le \epsilon$ is $\#\mathrm{grad} = O(n + \frac{\sqrt{n}L}{\epsilon^2})$.*

**Theorem 2 (Lower bound)**

*There exists a function $f$ satisfying average $L$-smoothness such that any linear-span first-order algorithm needs $\Omega(n + \frac{\sqrt{n}L}{\epsilon^2})$ stochastic gradient computations for finding an $\epsilon$-approximate solution.*

# Convergence Result and Lower Bound (online)

Average $L$-smooth: $\mathbb{E}_i[\|\nabla f_i(x) - \nabla f_i(y)\|^2] \le L^2\|x-y\|^2$

Bounded variance: $\mathbb{E}_i[\|\nabla f_i(x) - \nabla f(x)\|^2] \le \sigma^2$

## Theorem 3 (Optimal result of PAGE in online case)

*Suppose $f$ is average $L$-smooth and has bounded variance of stochastic gradient, choosing appropriate parameters, the number of stochastic gradient computations of PAGE for finding an $\epsilon$-approximate solution is* $\#\mathrm{grad} = O(b + \frac{\sqrt{b}L}{\epsilon^2})$, *where* $b = \min\{n, \frac{2\sigma^2}{\epsilon^2}\}$.

## Theorem 4 (Lower bound)

*There exists a function $f$ satisfying average $L$-smoothness and bounded variance of stochastic gradient such that any linear-span first-order algorithm needs* $\Omega(b + \frac{\sqrt{b}L}{\epsilon^2})$, *where* $b = \min\{n, \frac{\sigma^2}{\epsilon^2}\}$, *stochastic gradient computations for finding an $\epsilon$-approximate solution.*

# Better Convergence under PL Condition

**PL condition**: $\exists \mu > 0, \text{ such that } \|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*)$

If $f$ satisfies PL condition, PAGE will lead to faster linear convergence rates $O(\log \frac{1}{\epsilon})$ instead of sublinear rates $O(\frac{1}{\epsilon^2})$.

# Better Convergence under PL Condition

**PL condition**: $\exists \mu > 0, \text{ such that } \|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*)$

If $f$ satisfies PL condition, PAGE will lead to faster linear convergence rates $O(\log \frac{1}{\epsilon})$ instead of sublinear rates $O(\frac{1}{\epsilon^2})$.

**Theorem 5 (Switch to linear convergence under PL condition)**

*Under* **PL condition**, *PAGE with the same parameter setting can switch to the faster* **linear convergence** *results, i.e.,*

- *Finite-sum case:* $O(n + \frac{\sqrt{n}L}{\epsilon^2}) \longrightarrow \boldsymbol{O\Big((n + \frac{\sqrt{n}L}{\mu}) \log \frac{1}{\epsilon}\Big)}$
- *Online case:* $O(b + \frac{\sqrt{b}L}{\epsilon^2}) \longrightarrow \boldsymbol{O\Big((b + \frac{\sqrt{b}L}{\mu}) \log \frac{1}{\epsilon}\Big)}, \quad b = \min\{n, \frac{2\sigma^2}{\mu\epsilon}\}.$

# Experiments

Recall the update step of PAGE:

$$x^{t+1} = x^t - \eta g^t$$

$$g^{t+1} = \begin{cases} \frac{1}{b} \sum_{i \in I} \nabla f_i(x^{t+1}) & \text{with prob. } p_t \\ g^t + \frac{1}{b'} \sum_{i \in I'} (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) & \text{with prob. } 1 - p_t \end{cases}$$

PAGE is easy to implement via a small adjustment to vanilla SGD (i.e., $p = 1$ in PAGE), and enjoys a lower computational cost if $b' < b$.

In theory, PAGE can be better than SGD by a factor of $\frac{1}{\epsilon^2}$ or $\frac{\sigma}{\epsilon}$, where $\epsilon$ is the target error $\mathbb{E}[\|\nabla f(\hat{x}_T)\|] \leq \epsilon$.

# Experiments

Recall the update step of PAGE:

$$x^{t+1} = x^t - \eta g^t$$

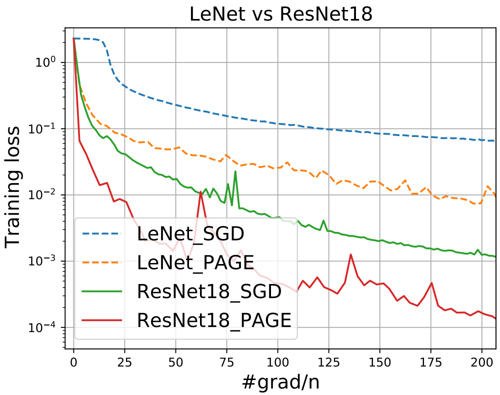
$$g^{t+1} = \begin{cases} \frac{1}{b}\sum_{i\in I}\nabla f_i(x^{t+1}) & \text{with prob. } p_t \\ g^t + \frac{1}{b'}\sum_{i\in I'}(\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) & \text{with prob. } 1-p_t \end{cases}$$

PAGE is easy to implement via a small adjustment to vanilla SGD (i.e., $p = 1$ in PAGE), and enjoys a lower computational cost if $b' < b$.

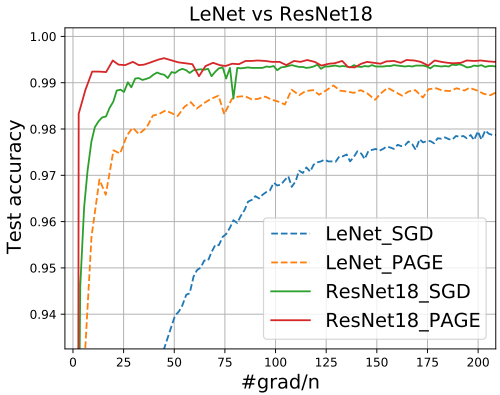
In theory, PAGE can be better than SGD by a factor of $\frac{1}{\epsilon^2}$ or $\frac{\sigma}{\epsilon}$, where $\epsilon$ is the target error $\mathbb{E}[\|\nabla f(\hat{x}_T)\|] \le \epsilon$.

In experiments, we compare our PAGE with SGD by running standard LeNet, VGG, ResNet models on MNIST and CIFAR-10 datasets.

# Experiments (MNIST)




(1a) Training loss

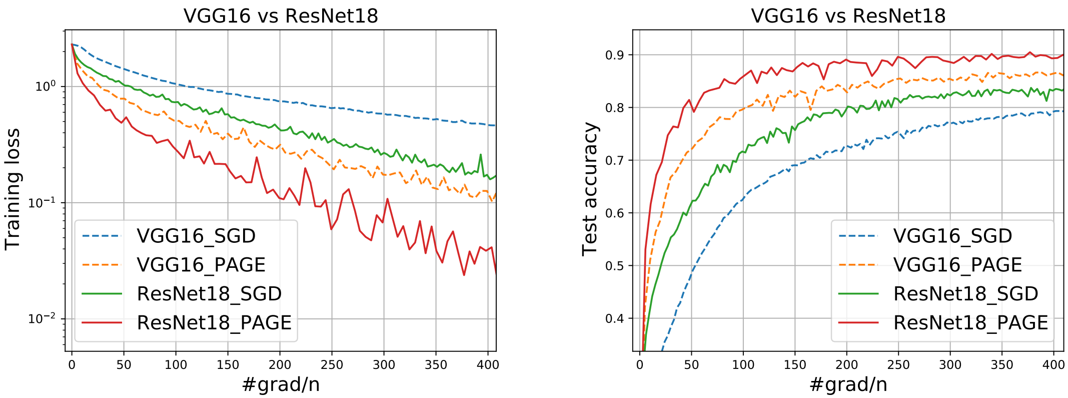



(1b) Test accuracy

Figure 1: PAGE vs. SGD with LeNet and ResNet18 on MNIST dataset

- PAGE converges *faster in training* and also gets *higher test accuracy*.

# Experiments (CIFAR-10)




(2a) Training loss (2b) Test accuracy

Figure 2: PAGE vs. SGD with VGG16 and ResNet18 on CIFAR-10 dataset

- Similarly, PAGE converges *faster in training* and gets *higher test accuracy* on CIFAR-10 dataset.

# Conclusion

- Propose a simple probabilistic gradient estimator called PAGE
- PAGE achieves optimal convergence rates for both nonconvex finite-sum and online problems
- Provide simple and clean convergence analysis, and tight lower bounds for validating the optimality
- PAGE can switch to a faster linear convergence under PL condition
- PAGE is easy to implement, converges faster in training and also achieves higher test accuracy than SGD

# Thanks!

Zhize Li

https://zhizeli.github.io