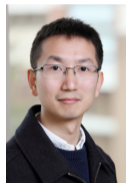


Stochastic Nested Variance Reduction for Nonconvex Optimization



Dongruo Zhou

Pan Xu

Quanquan Gu

Department of Computer Science
University of California, Los Angeles

- **Nonconvex finite-sum optimization:**

$$\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x}).$$

- ▶ f_i is L -smooth: $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2$.
- ▶ F has stochastic gradients with σ^2 -bounded variance: $\mathbb{E}_i \|\nabla f_i(\mathbf{x}) - \nabla F(\mathbf{x})\|_2^2 \leq \sigma^2$.
- ▶ (Optional) F is τ -gradient dominated (P-L condition): $F(\mathbf{x}) - \min_{\mathbf{y}} F(\mathbf{y}) \leq \tau \cdot \|\nabla F(\mathbf{x})\|_2^2$.

- **Nonconvex finite-sum optimization:**

$$\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x}).$$

- ▶ f_i is L -smooth: $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2$.
- ▶ F has stochastic gradients with σ^2 -bounded variance: $\mathbb{E}_i \|\nabla f_i(\mathbf{x}) - \nabla F(\mathbf{x})\|_2^2 \leq \sigma^2$.
- ▶ (Optional) F is τ -gradient dominated (P-L condition): $F(\mathbf{x}) - \min_{\mathbf{y}} F(\mathbf{y}) \leq \tau \cdot \|\nabla F(\mathbf{x})\|_2^2$.

- **Goals:**

- ▶ Find an ϵ -approximate first-order stationary point $\hat{\mathbf{x}}$ of F , such that

$$\|\nabla F(\hat{\mathbf{x}})\|_2 \leq \epsilon.$$

- ▶ If F is additionally τ -gradient dominated, find an ϵ -approximate global minimizer $\hat{\mathbf{x}}$, such that

$$F(\hat{\mathbf{x}}) - \inf_{\mathbf{x}} F(\mathbf{x}) \leq \epsilon.$$

Existing Algorithms & Convergence Results

- Gradient Descent (GD)
- Stochastic Gradient Descent (SGD)
- Stochastic Variance-reduced Gradient (SVRG) (Johnson and Zhang, 2013; Allen-Zhu and Hazan, 2016; Reddi et al., 2016)
- Stochastically Controlled Stochastic Gradient (SCSG) (Lei et al., 2017)

Existing Algorithms & Convergence Results

- Gradient Descent (GD)
- Stochastic Gradient Descent (SGD)
- Stochastic Variance-reduced Gradient (SVRG) (Johnson and Zhang, 2013; Allen-Zhu and Hazan, 2016; Reddi et al., 2016)
- Stochastically Controlled Stochastic Gradient (SCSG) (Lei et al., 2017)

Update rules: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \cdot \mathbf{v}_t$. $\tilde{\mathbf{x}}_t$ is the reference point of \mathbf{x}_t .

Algorithm	\mathbf{v}_t	Gradient Complexity
GD	$\nabla F(\mathbf{x}_t)$	$O(n\epsilon^{-2})$
SGD	$\nabla f_{\mathcal{I}_t}(\mathbf{x}_t)$	$O(\epsilon^{-4})$
SVRG	$\nabla f_{\mathcal{I}_t}(\mathbf{x}_t) - \nabla f_{\mathcal{I}_t}(\tilde{\mathbf{x}}_t) + \nabla F(\tilde{\mathbf{x}}_t)$	$O(n^{2/3}\epsilon^{-2})$
SCSG	$\nabla f_{\mathcal{I}_t}(\mathbf{x}_t) - \nabla f_{\mathcal{I}_t}(\tilde{\mathbf{x}}_t) + \nabla f_{\mathcal{I}_B}(\tilde{\mathbf{x}}_t)$	$O(n^{2/3}\epsilon^{-2} \wedge \epsilon^{-10/3})$

Gradient complexity: the number of stochastic gradient computations.

Stochastic Nested Variance Reduced Gradient Descent(SNVRG)

Algorithm 1 SNVRG-Epoch

- 1: **Input:** $\mathbf{x}_0, \eta, B, K, \{B_l\}, \{T_l\}$.
 - 2: Randomly pick \mathcal{I}_B with size B .
 - 3: $\mathbf{g}_0^{(0)} \leftarrow \nabla f_{\mathcal{I}_B}(\mathbf{x}_0), \mathbf{x}_0^{(0)} \leftarrow \mathbf{x}_0$
 - 4: $\mathbf{g}_0^{(l)} \leftarrow 0, \mathbf{x}_0^{(l)} \leftarrow \mathbf{x}_0, l \in [K]$.
 - 5: $\mathbf{v}_0 \leftarrow \sum_{l=0}^K \mathbf{g}_0^{(l)}, \mathbf{x}_1 \leftarrow \mathbf{x}_0 - \eta \cdot \mathbf{v}_0$.
 - 6: **for** $t = 1, \dots, \prod_{l=1}^K T_l - 1$ **do**
 - 7: Update $\{\mathbf{x}_t^{(l)}\}$ and $\{\mathbf{g}_t^{(l)}\}$.
 - 8: $\mathbf{v}_t \leftarrow \sum_{l=0}^K \mathbf{g}_t^{(l)}$.
 - 9: $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta \cdot \mathbf{v}_t$.
 - 10: **end for**
 - 11: $\mathbf{x}_{\text{out}} \leftarrow$ uniformly chosen from $\{\mathbf{x}_{0 \leq t < \prod_{l=1}^K T_l}\}$.
 - 12: **Output:** $[\mathbf{x}_{\text{out}}, \mathbf{x}_{\prod_{l=1}^K T_l}]$.
-

Algorithm 2 SNVRG

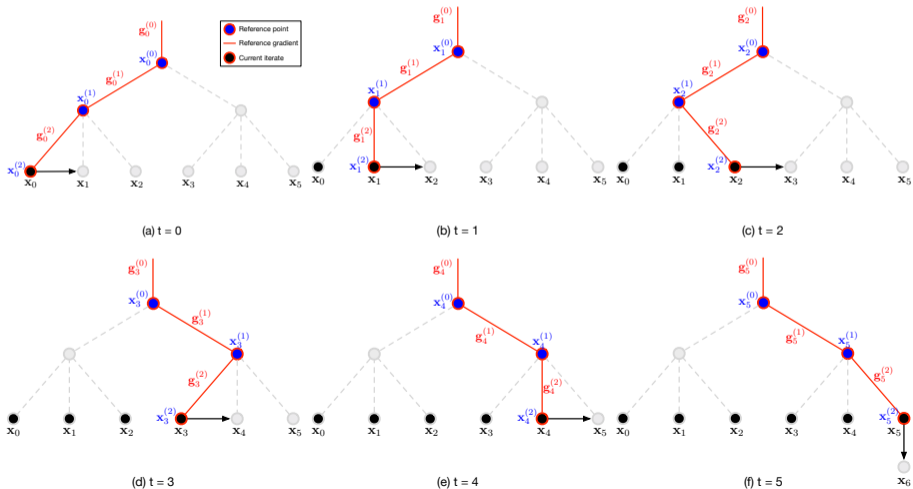
- 1: **Input:** $\mathbf{z}_0, \eta, B, K, \{B_l\}, \{T_l\}, S$.
 - 2: **for** $s = 1, \dots, S$ **do**
 - 3: $[\mathbf{y}_s, \mathbf{z}_s] \leftarrow$ SNVRG-Epoch
 $(\mathbf{z}_{s-1}, \eta, B, K, \{B_l\}, \{T_l\})$.
 - 4: **end for**
 - 5: **Output:** $\mathbf{y}_{\text{out}} \leftarrow$ uniformly chosen from $\{\mathbf{y}_s\}$.
-
-

Algorithm 3 SNVRG-PL

- 1: **Input:** $\mathbf{z}_0, \eta, B, K, \{B_l\}, \{T_l\}, S, U$.
 - 2: **for** $u = 1, \dots, U$ **do**
 - 3: $\mathbf{z}_u \leftarrow$ SNVRG
 $(\mathbf{z}_{u-1}, \eta, B, K, \{B_l\}, \{T_l\}, S)$.
 - 4: **end for**
 - 5: **Output:** $\mathbf{z}_{\text{out}} \leftarrow \mathbf{z}_U$.
-

Illustration of Update Rules

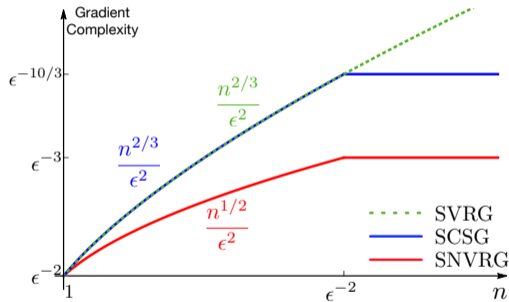
We take $K = 2$, $T_1 = 2$, $T_2 = 3$ as an example:



Gradient Complexity Comparison

Gradient complexity for finding an ϵ -approximate first-order stationary point:

Algorithm	Gradient Complexity
GD	$O(n\epsilon^{-2})$
SGD	$O(\epsilon^{-4})$
SVRG	$O(n^{2/3}\epsilon^{-2})$
SCSG	$O(n^{2/3}\epsilon^{-2} \wedge \epsilon^{-10/3})$
SNVRG (this paper)	$\tilde{O}(n^{1/2}\epsilon^{-2} \wedge \epsilon^{-3})$



- **SNVRG** is strictly better than SCSG by a factor of $\Omega(n^{1/6} \wedge \epsilon^{-1/3})$.
- A similar complexity has also been obtained in a concurrent work by [Fang et al., 2018](#).

Gradient Complexity Comparison under P-L Condition

Gradient complexity for finding an ϵ -approximate global minimizer:

Algorithm	Gradient Complexity
GD	$\tilde{O}(\tau n)$
SGD	$O(\epsilon^{-4})$
SVRG	$\tilde{O}(n + \tau n^{2/3})$
SCSG	$\tilde{O}(n \wedge \frac{\tau}{\epsilon} + \tau(n \wedge \frac{\tau}{\epsilon})^{2/3})$
SNVRG-PL (this paper)	$\tilde{O}(n \wedge \frac{\tau}{\epsilon} + \tau(n \wedge \frac{\tau}{\epsilon})^{1/2})$

- **SNVRG-PL** is strictly better than SCSG by a factor of $\Omega((n \wedge \tau \epsilon^{-1})^{1/6})$.

Thanks!

Poster session:

Wed Dec 5th 05:00 – 07:00 PM

@ Room 210 & 230 AB #44

UCLA

Stochastic Nested Variance Reduction for Nonconvex Optimization

Dongruo Zhou and Pan Xu and Quanquan Gu

Department of Computer Science, University of California, Los Angeles

Problem Setup and Preliminaries

- Optimization problem:** $\min_{x \in \mathbb{R}^n} F(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$, where f_i and F may be nonconvex.
- Assumptions:**
 - $F(x) \geq \rho^*$, $\forall x \in \mathbb{R}^n$.
 - $F(x) - F^* \leq \Delta$.
 - f_i is L -smooth, $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|$.
 - F is σ -variance bounded, $\mathbb{E}\|\nabla F(x) - \nabla F(x)\|^2 \leq \sigma^2$.
 - (Optional) F is μ -gradient dominated (P-L condition), $F(x) - F^* \leq \tau \|\nabla F(x)\|^2$.

Stochastic Nested Variance-Reduced Gradients

Algorithm 1 SNVRG-Epoch

- Input:** initial point x_0 , step size η , loop number K , base batch size B , batch parameters $\{B_i\}$, loop parameters $\{T_i\}$.
- Randomly pick $X_{i,0}$ with size B_i .
- $g_i^0 \leftarrow \nabla f_i(x_0)$, $x_i^0 \leftarrow x_0, g_i^0 \leftarrow 0, x_i^0 \leftarrow x_0, i \in [K]$.
- $y_i \leftarrow \sum_{j=0}^{T_i-1} g_i^j, x_{i,0} \leftarrow x_0, \forall i \in [K]$.
- for** $t = 1, \dots, \prod_{i=1}^K T_i - 1$ **do**
- Update $\{x_i^t\}$ and $\{g_i^t\}$.
- $y_i \leftarrow \sum_{j=0}^{T_i-1} g_i^j + g_i^t$.
- $x_{i,t+1} \leftarrow x_i^t + \eta y_i$.
- $x_{i,t+1} \leftarrow$ uniformly chosen from $\{x_{i,0}, \dots, x_{i,t}\}$.
- Output:** $\{x_{i,t}, \nabla f_i(x_{i,t})\}$.

Algorithm 2 SNVRG-PL

- Input:** initial point $x_0, \eta, B, K, \{B_i\}, \{T_i\}, S$, epoch number S .
- for** $s = 1, \dots, S$ **do**
- $\{y_i, x_i\} \leftarrow$ SNVRG-Epoch($x_0, \eta, B, K, \{B_i\}, \{T_i\}$).
- Output:** Uniformly chosen $x_{i,t}$ from $\{y_i\}$.

Algorithm 3 SNVRG-PL

- Input:** initial point $x_0, \eta, B, K, \{B_i\}, \{T_i\}, S$, epoch number S .
- for** $s = 1, \dots, S$ **do**
- $\{y_i, x_i\} \leftarrow$ SNVRG-Epoch($x_0, \eta, B, K, \{B_i\}, \{T_i\}$).
- Output:** $x_{i,t} \leftarrow x_i^t$.

Update rules for $\{x_i^t\}$ and $\{g_i^t\}$

- $r =$ the smallest number where r can be divided by $\prod_{i=1}^K T_i$.
- Update rules for reference points $\{x_i^t\}$:
 - x_i^0, \dots, x_i^{r-1} remain the same as x_i^0, \dots, x_i^0 .
 - x_i^r, \dots, x_i^{2r-1} remain the same as x_i^0, \dots, x_i^0 .
 - Update rules for reference gradients $\{g_i^t\}$:
 - g_i^0, \dots, g_i^{r-1} remain the same as g_i^0, \dots, g_i^0 .
 - For $r \leq t \leq K$, randomly pick up Z with size B_i , $g_i^t \leftarrow \nabla f_i(x_i^t) - \nabla f_i(x_i^{t-r})$.

Illustration of Update Rules

► **SNVRG-Epoch** with $K = 2, T_1 = 2, T_2 = 3$.

(a) $t = 0$ (b) $t = 1$ (c) $t = 2$ (d) $t = 3$ (e) $t = 4$ (f) $t = 5$ (g) $t = 6$ (h) $t = 7$

Theoretical Results

- Main Result:** Under Assumptions (1)-(4), with specific choice of parameters, **SNVRG-Epoch** outputs an ϵ -first-order stationary point $x_{i,t}$, i.e., $\mathbb{E}\|\nabla F(x_{i,t})\|^2 \leq \epsilon^2$, with

$$O\left(\log\left(\frac{\Delta}{\epsilon^2}\right) \left(\frac{\Delta}{\epsilon^2}\right)^{\frac{1}{\mu}} \left(\frac{\Delta}{\epsilon^2}\right)^{\frac{1}{\mu}} \left(\frac{\Delta}{\epsilon^2}\right)^{\frac{1}{\mu}}\right)$$
 stochastic gradient computations.
- Extension:** Under Assumptions (1)-(5), with specific choice of parameters, **SNVRG-PL** outputs an ϵ -accurate solution $x_{i,t}$, i.e., $\mathbb{E}[F(x_{i,t})] - F^* \leq \epsilon$, with

$$O\left(\log\left(\frac{\Delta}{\epsilon}\right) \left(\frac{\Delta}{\epsilon}\right)^{\frac{1}{\mu}} \left(\frac{\Delta}{\epsilon}\right)^{\frac{1}{\mu}} + \tau L \left(\frac{\Delta}{\epsilon}\right)^{\frac{1}{\mu}}\right)$$
 stochastic gradient computations.

Comparison with State-of-the-art

Table: Gradient complexities for finding an ϵ -first-order stationary point (nonconvex) or ϵ -accurate solution (gradient dominated)

Algorithm	nonconvex	gradient dominant
GD	$O(n^{3/2})$	$O(n^2)$
SGD	$O(n^{-2})$	$O(n^{-1})$
SVRG	$O(n^{3/4})$	$\tilde{O}(n + \tau n^{1/2})$
(Beck et al., 2018)		
SCSG	$O(n^{3/2} \Lambda, n^{3/4} \sigma^{-1/2})$	$\tilde{O}(n \Lambda + \tau(n \Lambda + \sigma^{1/2}))$
(Lei et al., 2022)		
SNVRG	$\tilde{O}(n^{3/4} \Lambda, n^{3/4} \sigma^{-1/2})$	$\tilde{O}(n \Lambda + \tau(n \Lambda + \sigma^{1/2}))$
(this paper)		

Figure: Gradient complexity comparison with SVRG and SCSG.

Numerical Experiments

- Baseline Algorithms:** SGD, SGD with momentum (Qian, 1999), ADAM (Kingma et al., 2014), SCSG (Lei et al., 2017)
- Datasets:** MNIST, CIFAR10, SVHN
- Training LeNet-5** (LeCun et al., 1999)

(a) Training loss (MNIST) (b) Training loss (CIFAR10) (c) Training loss (SVHN)

(d) Test error (MNIST) (e) Test error (CIFAR10) (f) Test error (SVHN)

Stochastic Nested Variance Reduction for Nonconvex Optimization