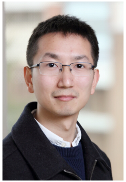
Stochastic Nested Variance Reduction for Nonconvex Optimization





Dongruo Zhou Pan Xu Quanquan Gu

Department of Computer Science
University of California, Los Angeles

Problem Setup

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

Problem Setup

- **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: $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2 \le L\|\mathbf{x}-\mathbf{y}\|_2$.
  - F has stochastic gradients with σ^2-bounded variance: $\mathbb{E}_i\|\nabla f_i(\mathbf{x}) - \nabla F(\mathbf{x})\|_2^2 \le \sigma^2$.
  - (Optional) F is τ-gradient dominated (P-L condition): $F(\mathbf{x}) - \min_{\mathbf{y}} F(\mathbf{y}) \le \tau \cdot \|\nabla F(\mathbf{x})\|_2^2$.
- **Goals:**
  - Find an ϵ-approximate first-order stationary point $\widehat{\mathbf{x}}$ of F, such that

$$\|\nabla F(\widehat{\mathbf{x}})\|_2 \le \epsilon.$$

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

$$F(\widehat{\mathbf{x}}) - \inf_{\mathbf{x}} F(\mathbf{x}) \le \epsilon.$$

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: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \cdot \mathbf{v}_t$. $\widetilde{\mathbf{x}}_t$ 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	$\nabla f_{\mathcal{I}_t}(\mathbf{x}_t) - \nabla f_{\mathcal{I}_t}(\widetilde{\mathbf{x}}_t) + \nabla F(\widetilde{\mathbf{x}}_t)$	$O(n^{2/3}\epsilon^{-2})$
SCSG	$\nabla f_{\mathcal{I}_t}(\mathbf{x}_t) - \nabla f_{\mathcal{I}_t}(\widetilde{\mathbf{x}}_t) + \nabla f_{\mathcal{I}_B}(\widetilde{\mathbf{x}}_t)$	$O(n^{2/3}\epsilon^{-2} \wedge \epsilon^{-10/3})$

Gradient complexity: the number of stochastic gradient computations.

Stochastic Nested Variance Reduced Gradient Descent(SNVRG)

Algorithm 1 SNVRG-Epoch

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

Algorithm 2 SNVRG

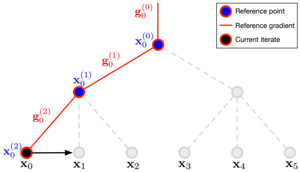
Input: $\mathbf{z}_0, \eta, B, K, \{B_l\}, \{T_l\}, S$.
for $s = 1, \ldots, S$ **do**
    $[\mathbf{y}_s, \mathbf{z}_s] \leftarrow$ SNVRG-Epoch $(\mathbf{z}_{s-1}, \eta, B, K, \{B_l\}, \{T_l\})$.
end for
Output: $\mathbf{y}_{\text{out}} \leftarrow$ uniformly chosen from $\{\mathbf{y}_s\}$.

Algorithm 3 SNVRG-PL

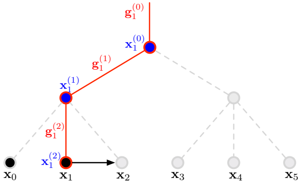
Input: $\mathbf{z}_0, \eta, B, K, \{B_l\}, \{T_l\}, S, U$.
for $u = 1, \ldots, U$ **do**
    $\mathbf{z}_u \leftarrow$ SNVRG $(\mathbf{z}_{u-1}, \eta, B, K, \{B_l\}, \{T_l\}, S)$.
end for
Output: $\mathbf{z}_{\text{out}} \leftarrow \mathbf{z}_U$.

Illustration of Update Rules

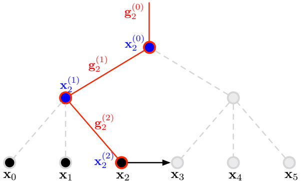
We take $K = 2$, $T_1 = 2$, $T_2 = 3$ as an example:



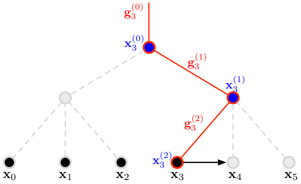
(a) t = 0



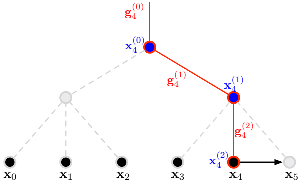
(b) t = 1



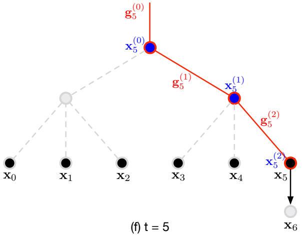
(c) t = 2



(d) t = 3



(e) t = 4

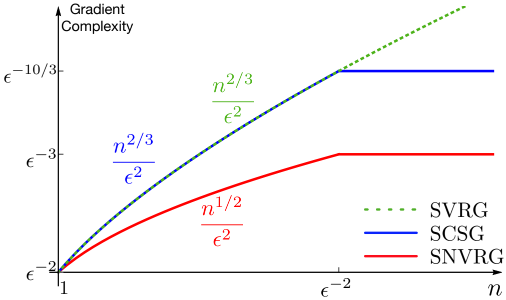


(f) t = 5

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)	$\widetilde{O}(n^{1/2}\epsilon^{-2} \wedge \epsilon^{-3})$



- **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	$\widetilde{O}(\tau n)$
SGD	$O(\epsilon^{-4})$
SVRG	$\widetilde{O}(n + \tau n^{2/3})$
SCSG	$\widetilde{O}(n \wedge \frac{\tau}{\epsilon} + \tau(n \wedge \frac{\tau}{\epsilon})^{2/3})$
SNVRG-PL (this paper)	$\widetilde{O}(n \wedge \frac{\tau}{\epsilon} + \tau(n \wedge \frac{\tau}{\epsilon})^{1/2})$

- **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**