An Optimal Hybrid Variance-Reduced Algorithm for
Stochastic Composite Nonconvex Optimization

Deyi Liu, Lam M. Nguyen, and Quoc Tran-Dinh

August 21, 2020

Abstract
In this note we propose a new variant of the hybrid variance-reduced proximal gradient method in [7] to solve a common stochastic composite nonconvex optimization problem under standard assumptions. We simply replace the independent unbiased estimator in our hybrid-SARAH estimator introduced in [7] by the stochastic gradient evaluated at the same sample, leading to the identical momentum-SARAH estimator introduced in [2]. This allows us to save one stochastic gradient per iteration compared to [7], and only requires two samples per iteration. Our algorithm is very simple and achieves optimal stochastic oracle complexity bound in terms of stochastic gradient evaluations (up to a constant factor). Our analysis is essentially inspired by [7], but we do not use two different step-sizes.

1 Problem Statement and Standard Assumptions
We consider the following stochastic composite and possibly nonconvex optimization problem:

$$\min_{x \in \mathbb{R}^p} \left\{ F(x) := \mathbb{E}_\xi [f_\xi(x)] + \psi(x) \right\}, \quad (1)$$

where $f_\xi(\cdot) : \mathbb{R}^p \times \Omega \to \mathbb{R}$ is a stochastic function defined, such that for each $x \in \mathbb{R}^p$, $f_\xi(x)$ is a random variable in a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, while for each realization $\xi \in \Omega$, $f_\xi(\cdot)$ is differentiable on $\mathbb{R}^p$; and $f(x) := \mathbb{E}_\xi [f_\xi(x)]$ is the expectation of the random function $f_\xi(x)$ over $\xi$ on $\Omega$: $\psi : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is a proper, closed, and convex function.

Our algorithm developed in this note relies on the following fundamental assumptions:

**Assumption 1.1.** The objective functions $f$ and $\psi$ of (1) satisfies the following conditions:

(a) **(Convexity of $\psi$)** $\psi : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is proper, closed, and convex. In addition, $\text{dom}(F) := \text{dom}(f) \cap \text{dom}(\psi) \neq \emptyset$.

(b) **(Boundedness from below)** There exists a finite lower bound

$$F^* := \inf_{x \in \mathbb{R}^p} \left\{ F(x) := f(x) + \psi(x) \right\} > -\infty. \quad (2)$$

(c) **(L-average smoothness)** The expectation function $f(\cdot)$ is $L$-smooth on $\text{dom}(F)$, i.e., there exists $L \in (0, +\infty)$ such that

$$\mathbb{E}_\xi \left[ \|\nabla f_\xi(x) - \nabla f_\xi(y)\|^2 \right] \leq L^2 \|x - y\|^2, \quad \forall x, y \in \text{dom}(F). \quad (3)$$

(d) **(Bounded variance)** There exists $\sigma \in [0, +\infty)$ such that

$$\mathbb{E}_\xi \left[ \|\nabla f_\xi(x) - \nabla f(x)\|^2 \right] \leq \sigma^2, \quad \forall x \in \text{dom}(F). \quad (4)$$

These assumptions are very standard in stochastic optimization and required for various gradient-based methods. Unlike [2], we do not impose a bounded gradient assumption, i.e., $\|\nabla f(x)\| \leq G$ for all $x \in \mathbb{R}^p$. Algorithm 1 below has a single loop and achieves optimal oracle complexity bound since it matches the lower bound complexity in [1] up to a constant factor.
2 Hybrid Variance-Reduced Proximal Gradient Algorithm

We first propose a new variant of [7] Algorithm 1 for solving (1) and then analyze its convergence and oracle complexity.

2.1 Main result: Algorithm and its convergence

We propose a novel hybrid variance-reduced proximal gradient method to solve (1) under standard assumptions (i.e., Assumption [1.1]) as described in Algorithm [1.1].

Algorithm 1 (Hybrid Variance-Reduced Proximal Gradient Algorithm)

1: Initialization: An arbitrarily initial point \(x_0 \in \text{dom}(F)\).
2: Choose an initial batch size \(b \geq 1\), \(\beta \in (0, 1)\), and \(k > 0\) as in Theorem 2.1 below.
3: Generate an unbiased estimator \(v_0 := \frac{1}{b} \sum_{i \in \hat{B}} \nabla f_{\xi_i}(x_0)\) at \(x_0\) using a mini-batch \(\hat{B}\).
4: Update \(x_1 := \text{prox}_{\eta v_0}(x_0 - \eta_0 v_0)\).
5: For \(t := 1, \ldots, T\) do
6: Generate a proper sample \(\xi_t\) (single sample or mini-batch).
7: Evaluate \(v_t\) and update
   \[
   \begin{align*}
   v_t &:= \nabla f_{\xi_t}(x_t) + (1 - \beta) [v_{t-1} - \nabla f_{\xi_t}(x_{t-1})] \\
x_{t+1} &:= \text{prox}_{\eta v_t}(x_t - \eta v_t).
   \end{align*}
   \]
8: EndFor
9: Choose \(\pi_T\) uniformly from \(\{x_0, x_1, \ldots, x_T\}\).

Compared to [7] Algorithm 1, the new algorithm, Algorithm [1.1] has two major differences. First, it uses a new estimator \(v_t\) adopted from [2]. This estimator can also be viewed as a variant of the hybrid SARAH estimator in [7] by using the same sample \(\xi_t\) for \(\nabla f_{\xi_t}(x_t)\). That is

Hybrid SARAH [7]:

\[v^h_t := (1 - \beta)[v_{t-1} + \nabla f_{\xi_t}(x_t) - \nabla f_{\xi_t}(x_{t-1})] + \beta \nabla f_{\xi_t}(x_t), \ \xi_t \neq \xi_i,\]

Momentum SARAH [2]:

\[v_t := (1 - \beta)[v_{t-1} + \nabla f_{\xi_t}(x_t) - \nabla f_{\xi_t}(x_{t-1})] + \beta \nabla f_{\xi_t}(x_t), \ \xi_t = \xi_i.\]

Second, it does not require an extra damped step-size \(\gamma\) as in [7], making Algorithm [1.1] simpler than the one in [7].

To analyze Algorithm [1.1] as usual, we define the following gradient mapping of (1):

\[G_\eta(x) := \frac{1}{\eta} \left( x - \text{prox}_{\eta v_t}(x - \eta \nabla f(x)) \right),\]

where \(\eta > 0\) is any given step-size. It is obvious to show that \(x^* \in \text{dom}(F)\) is a stationary point of (1), i.e., \(0 \in \nabla f(x^*) + \partial \psi(x^*)\) if and only if \(G_\eta(x^*) = 0\). We will show that for any \(\varepsilon > 0\), Algorithm [1.1] can find \(\pi_T\) such that \(E \left[ \|G_\eta(\pi_T)\|^2 \right] \leq \varepsilon^2\), which means that \(\pi_T\) is an \(\varepsilon\)-approximate stationary point of (1), where the expectation is taken over all the present randomness.

The following theorem establishes convergence of Algorithm [1.1] and provides oracle complexity.

**Theorem 2.1.** Under Assumption [1.1], assume that \(\eta \in (0, \frac{1}{2L})\) is a given step-size and \(0 < \frac{2\beta^2}{1 - 2\beta} \leq \beta < 1\). Let \(\{x_t\}_{t=0}^T\) be generated by Algorithm [1.1]. Then, we have

\[
\frac{1}{T+1} \sum_{t=0}^T E \left[ \|G_\eta(x_t)\|^2 \right] \leq \frac{2\left[ F(x_0) - F^* \right]}{\eta(T+1)} + \frac{E \left[ \|v_0 - \nabla f(x_0)\|^2 \right]}{\beta(T+1)} + 2\beta \sigma^2.
\]

In particular, if we choose \(\eta := \frac{1}{2\sqrt{(T+1)/\sigma}}\), \(\beta := \frac{1}{(T+1)^{1/3}}\), and \(b := \left[ (T+1)^{1/3} \right] \geq 1\), then the output \(\pi_T\) of Algorithm [1.1] satisfies

\[
E \left[ \|G_\eta(\pi_T)\|^2 \right] \leq \frac{4L \left[ F(x_0) - F^* \right]}{(T+1)^{2/3}} + 4\sigma^2.
\]
Consequently, for any tolerance $\varepsilon > 0$, the total number of stochastic gradient evaluations in Algorithm 2 to achieve $\mathcal{T}_T$ such that $E \left[ \left\| G_0(\mathcal{T}_T) \right\|^2 \right] \leq \varepsilon^2$ is at most $T_{\nabla f} := \left[ \frac{\Delta_0^{1/2}}{2\varepsilon} + \frac{\Delta_0^{3/2}}{\varepsilon^2} \right]$, where $\Delta_0 := 4 \left[ LF(x_0) - F^* \right] + \sigma^2$.

Theorem 2.1 shows that the oracle complexity of Algorithm 1 is $O \left( \frac{\Delta_0^{1/2}}{\varepsilon} + \frac{\Delta_0^{3/2}}{\varepsilon^2} \right)$ as in [7], where $\Delta_0 := 4 \left( LF(x_0) - F^* \right) + \sigma^2$. This complexity bound in fact matches the lower bound one in [1] up to a constant factor under the same assumptions as in Assumption 1.1. Hence, we conclude that Algorithm 1 is optimal.

2.2 Convergence Analysis

Let us denote by $\mathcal{F}_t := \sigma(\xi_0, \xi_1, \ldots, \xi_t)$ the σ-filed generated by $\{\xi_0, \xi_1, \ldots, \xi_t\}$. We also denote by $E[\cdot]$ the full expectation over the history $\mathcal{F}_t$. The following lemma establishes a key estimate for our convergence analysis. We emphasize that Lemma 2.1 is self-contained and can be applied to other types of estimators, e.g., Hessian, and other problems.

**Lemma 2.1.** Let $v_t$ be computed by (9) for $\beta \in (0, 1)$. Then, under Assumption 1.1, we have

$$E_{\xi_t} \left[ \|v_t - \nabla f(x_t)\|^2 \right] \leq (1 - \beta)^2 \|v_{t-1} - \nabla f(x_{t-1})\|^2 + 2(1 - \beta)^2 L^2 \|x_t - x_{t-1}\|^2 + 2\beta^2 \sigma^2. \quad (9)$$

Therefore, by induction, we have

$$E \left[ \|v_t - \nabla f(x_t)\|^2 \right] \leq (1 - \beta)^2 E \left[ \|v_0 - \nabla f(x_0)\|^2 \right] + 2\beta \sigma^2 + 2 L^2 \sum_{i=0}^{t-1} (1 - \beta)^{2\left( t-i \right)} E \left[ \|x_{i+1} - x_i\|^2 \right]. \quad (10)$$

**Proof.** Let us denote

$$a_t := (1 - \beta) [\nabla f_{\xi_t}(x_t) - \nabla f(x_t) - \nabla f_{\xi_t}(x_{t-1}) + \nabla f(x_{t-1})] \quad \text{and} \quad b_t := \beta [\nabla f_{\xi_t}(x_t) - \nabla f(x_t)].$$

Since $E_{\xi_t} [a_t] = E_{\xi_t} [b_t] = 0$, and (9), we can derive (10) as follows:

$$E_{\xi_t} \left[ \|v_t - \nabla f(x_t)\|^2 \right] = E_{\xi_t} \left[ \|\nabla f_{\xi_t}(x_t) + (1 - \beta)(v_{t-1} - \nabla f_{\xi_t}(x_{t-1})) - \nabla f(x_t)\|^2 \right]$$

$$= E_{\xi_t} \left[ \| (1 - \beta) \|v_{t-1} - \nabla f(x_{t-1})\| + a_t + b_t \|^2 \right]$$

$$= (1 - \beta)^2 \|v_{t-1} - \nabla f(x_{t-1})\|^2 + E_{\xi_t} \left[ \|a_t + b_t\|^2 \right]$$

$$\leq (1 - \beta)^2 \|v_{t-1} - \nabla f(x_{t-1})\|^2 + 2 E_{\xi_t} \left[ \|a_t\|^2 \right] + 2 E_{\xi_t} \left[ \|b_t\|^2 \right]$$

$$\leq (1 - \beta)^2 \|v_{t-1} - \nabla f(x_{t-1})\|^2 + 2(1 - \beta)^2 E_{\xi_t} \left[ \|\nabla f_{\xi_t}(x_t) - \nabla f_{\xi_t}(x_{t-1})\|^2 \right]$$

$$+ 2\beta^2 E_{\xi_t} \left[ \|\nabla f_{\xi_t}(x_t) - \nabla f(x_t)\|^2 \right]$$

$$\leq (1 - \beta)^2 \|v_{t-1} - \nabla f(x_{t-1})\|^2 + 2(1 - \beta)^2 L^2 \|x_t - x_{t-1}\|^2 + 2\beta^2 \sigma^2.$$
Lemma 2.2. Let \{x_t\} be generated by Algorithm 1 for solving (11) and G_\eta be defined by (6). Then, under Assumption 1.1, we have

\[
\mathbb{E}[F(x_{t+1}) - F^*] \leq \mathbb{E}[F(x_t) - F^*] - \left( \frac{1}{2\eta} - \frac{L}{2} \right) \mathbb{E} \left[ \|x_{t+1} - x_t\|^2 \right] - \frac{\eta}{2} \mathbb{E} \left[ \|G_\eta(x_t)\|^2 \right] + \frac{\eta}{2} \mathbb{E} \left[ \|\nabla f(x_t) - v_t\|^2 \right].
\]

(11)

Proof. Let us denote by \(\bar{x}_t := \text{prox}_{\eta \psi}(x_t - \eta \nabla f(x_t))\). From the optimality condition of this proximal operator, we have

\[
\langle \nabla f(x_t), \bar{x}_t - x_t \rangle + \frac{1}{2\eta}\|\bar{x}_t - x_t\|^2 + \psi(\bar{x}_t) \leq \psi(x_t) - \frac{1}{2\eta}\|x_t - \bar{x}_t\|^2.
\]

Similarly, from \(x_{t+1} = \text{prox}_{\eta \psi}(x_t - \eta v_t)\), we also have

\[
\langle v_t, x_{t+1} - x_t \rangle + \frac{1}{2\eta}\|x_{t+1} - x_t\|^2 + \psi(x_{t+1}) \leq \langle v_t, \bar{x}_t - x_t \rangle + \frac{1}{2\eta}\|\bar{x}_t - x_t\|^2 + \psi(\bar{x}_t) - \frac{1}{2\eta}\|\bar{x}_t - x_{t+1}\|^2.
\]

Combining the last two inequalities, we can show that

\[
\psi(x_{t+1}) + \frac{1}{2\eta}\|x_{t+1} - x_t\|^2 \leq \psi(x_t) - \frac{\eta}{2}\|G_\eta(x_t)\|^2 - \frac{\eta}{2\eta}\|\bar{x}_t - x_{t+1}\|^2 + \langle v_t, \bar{x}_t - x_{t+1} \rangle - \langle \nabla f(x_t), \bar{x}_t - x_t \rangle.
\]

(12)

By the Cauchy-Schwarz inequality, for any \(\eta > 0\), we easily get

\[
\langle \nabla f(x_t) - v_t, x_{t+1} - \bar{x}_t \rangle \leq \frac{\eta}{2}\|\nabla f(x_t) - v_t\|^2 + \frac{1}{2\eta}\|x_{t+1} - \bar{x}_t\|^2.
\]

(13)

Finally, using the L-average smoothness of f, we can derive

\[
f(x_{t+1}) + \psi(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 + \psi(x_{t+1})
\]

\[
= f(x_t) - \left( \frac{1}{2\eta} - \frac{L}{2} \right)\|x_{t+1} - x_t\|^2 + \langle \nabla f(x_t), x_{t+1} - x_t \rangle
\]

\[
+ \psi(x_{t+1}) + \frac{1}{2\eta}\|x_{t+1} - x_t\|^2 \overset{(12)}{\leq} f(x_t) - \left( \frac{1}{2\eta} - \frac{L}{2} \right)\|x_{t+1} - x_t\|^2 + \psi(x_t) + \langle \nabla f(x_t) - v_t, x_{t+1} - \bar{x}_t \rangle
\]

\[
- \frac{\eta}{2}\|G_\eta(x_t)\|^2 - \frac{1}{2\eta}\|\bar{x}_t - x_{t+1}\|^2 \overset{(13)}{\leq} f(x_t) + \psi(x_t) - \left( \frac{1}{2\eta} - \frac{L}{2} \right)\|x_{t+1} - x_t\|^2 + \frac{\eta}{2}\|\nabla f(x_t) - v_t\|^2 - \frac{\eta}{2\eta}\|G_\eta(x_t)\|^2.
\]

Taking the full expectation of both sides of the last inequality and noting that \(F = f + \psi\), we obtain (11). \(\square\)

Now, we are ready to prove our main result, Theorem 2.1 above.

The proof of Theorem 2.1 First, summing up (10) from \(t := 0\) to \(t := T\), we get

\[
\sum_{t=0}^T \mathbb{E} \left[ \|v_t - \nabla f(x_t)\|^2 \right] \leq \sum_{t=0}^T (1 - \beta)^{2t} \|v_0 - \nabla f(x_0)\|^2 + 2(T + 1)\beta \sigma^2
\]

\[
+ 2L^2 \sum_{t=0}^T \sum_{i=0}^{t-1} (1 - \beta)^{2(t-i)} \mathbb{E} \left[ \|x_{i+1} - x_i\|^2 \right] \leq \frac{1}{\beta} \|v_0 - \nabla f(x_0)\|^2 + 2(T + 1)\beta \sigma^2
\]

\[
+ 2L^2 \sum_{t=0}^{T-1} \sum_{i=0}^{t+1} (1 - \beta)^{2(t-i)} \mathbb{E} \left[ \|x_{i+1} - x_i\|^2 \right] \leq \frac{1}{\beta} \|v_0 - \nabla f(x_0)\|^2 + 2(T + 1)\beta \sigma^2
\]

\[
+ 2L^2 \sum_{i=0}^{T-1} \frac{1}{\beta} \mathbb{E} \left[ \|x_{i+1} - x_i\|^2 \right].
\]
Next, summing up (11) from $t := 0$ to $t := T$, we obtain

$$\mathbb{E}[F(x_{T+1}) - F^*] \leq [F(x_0) - F^*] - \frac{\eta}{2} \sum_{t=0}^{T} \mathbb{E} \left[ \|G_\eta(x_t)\|^2 \right] - \sum_{t=0}^{T} \left( \frac{1}{2\eta} - \frac{\beta}{2} \right) \mathbb{E} \left[ \|x_{t+1} - x_t\|^2 \right]$$

$$+ \frac{\eta}{2} \sum_{t=0}^{T} \mathbb{E} \left[ \|v_t - \nabla f(x_t)\|^2 \right]$$

$$\leq [F(x_0) - F^*] - \frac{\eta}{2} \sum_{t=0}^{T} \mathbb{E} \left[ \|G_\eta(x_t)\|^2 \right] - \sum_{t=0}^{T} \left( \frac{1}{2\eta} - \frac{\beta}{2} \right) \mathbb{E} \left[ \|x_{t+1} - x_t\|^2 \right]$$

$$+ \frac{\eta}{2\eta} \mathbb{E} \left[ \|v_0 - \nabla f(x_0)\|^2 \right] + \sum_{t=0}^{T-1} \frac{L^2\eta}{2} \mathbb{E} \left[ \|x_{t+1} - x_t\|^2 \right] + (T + 1)\eta\beta\sigma^2.$$

Since $\eta \in (0, \frac{1}{2L_\eta^2}]$, we have $0 < \frac{L^2\eta^2}{1 - L_\eta^2} < 1$. Suppose $\frac{1}{2\eta} - \frac{\beta}{2} \geq \frac{L^2\eta^2}{1 - L_\eta^2}$, i.e., $\beta \geq \frac{2L^2\eta^2}{1 - L_\eta^2}$, we have

$$\mathbb{E}[F(x_{T+1}) - F^*] \leq [F(x_0) - F^*] - \frac{\eta}{2} \sum_{t=0}^{T} \mathbb{E} \left[ \|G_\eta(x_t)\|^2 \right] + \frac{\eta}{2\beta} \mathbb{E} \left[ \|v_0 - \nabla f(x_0)\|^2 \right] + (T + 1)\eta\beta\sigma^2,$$

which leads to (7).

Now, if we choose $\eta := \frac{1}{2L(T+1)^{1/3}}$ and $\beta := \frac{1}{(T+1)^{2/3}}$, then we can verify that $\beta \geq \frac{2L^2\eta^2}{1 - L_\eta^2}$. Moreover, (7) becomes

$$\frac{1}{T + 1} \sum_{t=0}^{T} \mathbb{E} \left[ \|G_\eta(x_t)\|^2 \right] \leq \frac{4L}{(T + 1)^{2/3}} [F(x_0) - F^*] + \frac{2\sigma^2}{(T + 1)^{1/3}} + \mathbb{E} \left[ \|v_0 - \nabla f(x_0)\|^2 \right].$$

By Step 3 of Algorithm 1 and the choice $\tilde{b} := \left( \frac{L}{T+1} \right)^{1/3}$, we have $\mathbb{E} \left[ \|v_0 - \nabla f(x_0)\|^2 \right] \leq \frac{\sigma^2}{b} \leq \frac{2\sigma^2}{(T+1)^{1/3}}$. Substituting this bound into the previous one and using $\mathbb{E} \left[ \|G_\eta(x_T)\|^2 \right] = \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E} \left[ \|G_\eta(x_t)\|^2 \right]$, we obtain (8).

Finally, from (8), to guarantee $\mathbb{E} \left[ \|G_\eta(x_T)\|^2 \right] \leq \varepsilon^2$, we have $T + 1 \geq \frac{\Delta_0^3}{4\varepsilon^2}$. Therefore, the number of stochastic gradient evaluation is $\mathcal{T}_{\mathcal{G}} = \tilde{b} + 2T = \frac{\Delta_0^{1/2}}{2\varepsilon} + \frac{2\Delta_0^{3/2}}{\varepsilon^2}$. Rounding it, we obtain $\mathcal{T}_{\mathcal{G}} = \left[ \frac{\Delta_0^{1/2}}{2\varepsilon} + \frac{2\Delta_0^{3/2}}{\varepsilon^2} \right]$. \qed

### 3 Concluding Remarks and Outlook

Theorem 2.1 only analyzes a simple variant of Algorithm 1 with constant step-size $\eta = \mathcal{O} \left( \frac{1}{\sqrt{T}} \right)$ and constant weight $\beta = \mathcal{O} \left( \frac{1}{\sqrt{T}} \right)$. It also uses a large initial mini-batch of size $\bar{b} = \mathcal{O} \left( T^{1/3} \right)$. Compared to SARAH-based methods, e.g., in [3, 4, 5], Algorithm 1 is simpler since it is single-loop. At each iteration, it uses only two samples compared to three ones in [7]. We remark that the convergence of Algorithm 1 can be established by means of Lyapunov function as in [2].

The result of this note can be extended into different directions:

- We can also adapt our analysis to mini-batch, adaptive step-size $\eta_t$, and adaptive weight $\beta_t$ variants as in [6]. If we use adaptive weight $\beta_t$ as in [6], then we can remove the initial batch $\tilde{b}$ at Step 3 of Algorithm 1. However, the convergence rate in Theorem 2.1 will be $\mathcal{O} \left( \frac{\log(T)}{T^{1/3}} \right)$ instead of $\mathcal{O} \left( \frac{1}{T^{1/3}} \right)$. The rate $\mathcal{O} \left( \frac{\log(T)}{T^{1/3}} \right)$ matches the result of [2] without bounded gradient assumption.

- Our results, especially Lemma 2.3, can be applied to develop stochastic algorithms for solving other optimization problems such as compositional nonconvex optimization, minimax problems, and reinforcement learning.

- The idea here can also be extended to develop second-order methods such as sub-sampled and sketching Newton or cubic regularization-based methods.

It is also interesting to incorporate this idea with adaptive schemes as done in [2] by developing different strategies such as curvature aid or quasi-Newton methods.
References

1. Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. arXiv preprint arXiv:1912.02365, 2019.

2. A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex SGD. In Advances in Neural Information Processing Systems, pages 15210–15219, 2019.

3. C. Fang, C. J. Li, Z. Lin, and T. Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path integrated differential estimator. In Advances in Neural Information Processing Systems, pages 689–699, 2018.

4. L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. ICML, 2017.

5. H. N. Pham, M. L. Nguyen, T. D. Phan, and Q. Tran-Dinh. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. J. Mach. Learn. Res., 21:1–48, 2020.

6. Q. Tran-Dinh, D. Liu, and L. M. Nguyen. Hybrid variance-reduced SGD algorithms for nonconvex-concave minimax problems. Tech. Report STOR.05.20, UNC-Chapel Hill (arXiv preprint arXiv:2006.15266), 2020.

7. Q. Tran-Dinh, N. H. Pham, D. T. Phan, and L. M. Nguyen. A hybrid stochastic optimization framework for stochastic composite nonconvex optimization. arXiv preprint arXiv:1907.03793, pages 1–49, 2019.

Authors’ information:
Deyi Liu and Quoc Tran-Dinh*
Department of Statistics and Operations Research
The University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
Email: deyi@live.unc.edu, quoctd@email.unc.edu
*Corresponding author.

Lam M. Nguyen, IBM Research, Thomas J. Watson Research Center, NY10598
Email: lamnguyen.mltd@ibm.com