# A Note on Fully Non-convex Composite Optimization

May 24, 2018

**Abstract**

In this paper, we put forward SMART+, a randomized SVRG for fully non-convex non-smooth composite optimization. SMART+ has expected stochastic first-order oracle (*SFO*) complexity of $\mathcal{O}\left(n+\frac{n^{\frac{2}{3}}}{\epsilon}\right)$ for obtaining an $\epsilon$-accurate[1] solution. Our algorithm is able to handle the case where the target and the regularizer are both non-convex, and the regularizer can still be non-smooth. An online variant of our algorithm has expected *SFO* complexity of $\mathcal{O}\left(\epsilon^{-\frac{5}{3}}\right)$.

## 1 Introduction

We consider the following fully non-convex composite optimization problem.

$$\min_{x\in\mathcal{H}} F(x) := f(x) + g(x)$$

Here $f(x)=\frac{1}{n}\sum_{i=1}^{n} f_i(x)$, and each $f_i(x)$ is $C^1$ and has $L$-Lipschitz continuous gradient. We assume that $g(x)$ is a possibly non-convex regularizer, but proper and lower-semicontinuous.

This form of composite optimization problem and its variants has been widely studied in both computer science literature and mathematics literature. It is a fundamental problem for many machine learning applications, such as SVM, Lasso, neural network, etc. There has been abundant research when $f(x)$ and $g(x)$ are convex [cites], and we won't elaborate on them. A notable result is achieved by Katyusha. It has *SFO* complexity of $\mathcal{O}\left(n+\frac{\sqrt{n}}{\epsilon}\right)$ to achieve a solution satisfying $\mathbb{E}(F(\hat{x})-F(x^*))\leq\epsilon$, which matches the lower bound for this problem.

However, recently there have been more and more researches considering over non-convex composite optimization problems, which is intrinsically harder than their convex counterparts. Classical problem settings for this line of research is that each $f_i(x)$ is non-convex but $C^1$ and has $L$-Lipschitz gradient, and $g(x)$ is $C^0$ and convex. Representative methods for this problem are ProxSVRG and Natasha1.0, both achieving a complexity upper bound of $\mathcal{O}\left(n+\frac{n^{\frac{2}{3}}}{\epsilon}\right)$ to get an $\epsilon$-accurate result .

Online variant for the non-convex composite optimization problem usually make the assumption (explicitly or implicitly) that $n$ is quite large and even infinite, i.e., $n\gg\mathcal{O}(\frac{1}{\epsilon})$. Thus the complexity results of the algorithms designed for this problem should be independent of $n$. The state-of-the-art

[1] See preliminaries for definition.

established for this problem is the upper complexity bound $\mathcal{O}(\epsilon^{-\frac{5}{3}})$ achieved by [SCSG, Natasha1.5, ProxSVRG+].

However, in this work, we study a more general problem, where we assume that $f(x)$ is non-convex and $g(x)$ is $\delta$-non-convex (see Definition 4) and non-smooth. We dub it fully non-convex composite optimization. Although many researches have been devoted to the composite optimization with convex regularizers, there have been scarce literature covering this fully non-convex composite optimization, where the regularizer can be non-convex as well. One line of research is the non-monotone accelerated proximal gradient (APG) [Wang et al.] and non-convex APG [Yao et al.], both designed for fully non-convex optimization. [Li et al.] proved that these methods achieves linear and sub-linear rates under the KL condition. These two methods tried to use momentum in non-convex composite optimization to accelerate the progress, but their method does not guarantee better performance than gradient descent in general, let alone against the stochasticity. And the momentum trick does not improve the complexity essentially.

Another line of research on fully non-convex composite optimization is established by the proximal alternating linearized minimization (PALM) [Attouch et al.] and Asynchronous PALM [Davis et al.], both using the Gauss-Seidel method to solve composite optimization problems. However, the APALM requires a special form of composite function, and it concentrates on asynchronous settings. Both PALM and APALM need KL condition to achieve linear or sub-linear convergence rates.

However, in SMART [Aleksandr, et. al.], a randomization mechanism is introduced to solving this group of problem, and decent convergence results are established for general non-convex functions. It also uses the Gauss-Seidel method to solve a more complex composite optimization problem.

In light of their work, we introduce the randomization mechanism to the ProxSVRG. And we propose SMART+, which is targeted for fully nonconvex composite optimization. Our SMART+ only needs slight changes (only one to two lines of codes) over the ProxSVRG method, and achieves *SFO* complexity of $\mathcal{O}\left(n + \frac{n^{\frac{2}{3}}}{\epsilon}\right)$ despite the non-convex regularizer. We summarize the problem settings into the following three cases.

1. $f(x)$ is convex, $C^1$ and has $L$-Lipschitz smooth gradient, $g(x)$ is convex, proper and lower semicontinuous.

2. $f(x)$ is non-convex, $C^1$ and has $L$-Lipschitz smooth gradient, $g(x)$ is convex, proper and lower semicontinuous.

3. $f(x)$ is non-convex, $C^1$ and has $L$-Lipschitz smooth gradient, $g(x)$ is $\delta$-non-convex, proper and lower semicontinuous.

Notice that the problem hardness increases with its index. And we have the following complexity results.

| Settings | Offline | Online |
|---|---|---|
| (1) | $\mathcal{O}\left(n + \frac{\sqrt{n}}{\epsilon}\right)$ [Katyusha] | $\mathcal{O}(\epsilon^{-\frac{3}{2}})$ |
| (2) | $\mathcal{O}\left(n + \frac{n^{\frac{2}{3}}}{\epsilon}\right)$ [Natasha1.0, ProxSVRG] | $\mathcal{O}(\epsilon^{-\frac{5}{3}})$[SCSG, Natasha1.5] |
| (3) | $\mathcal{O}\left(n + \frac{n^{\frac{2}{3}}}{\epsilon}\right)$ [this work] | $\mathcal{O}(\epsilon^{-\frac{5}{3}})$[this work] |

The rest of the paper is organized as follows. In Section 2, we introduce some basic notions on composite optimization and non-convex optimization. Then we present our algorithm and its convergence results in Section 3 and Section 4 respectively. In Section 5, we consider an online variant of our algorithm, SMART+ online, and give its convergence results. We conclude our work in Section 6.

# 2 Preliminaries

We first introduce some basic definitions from optimization literatures.

**Definition 1.** (*L-Lipschitz Continuous*) The function $f(x)$ has $L$-Lipschitz continuous gradient if we have

$$\forall x, y \in \mathcal{H}, \qquad \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

**Definition 2.** (*Proper Function*) The function $g(x)$ is proper if its effective domain is non-empty and its value never attains $-\infty$.

**Definition 3.** (*Lower Semi-continuity*) The function $g(x)$ is lower semi-continuous if $\forall x_0 \in \mathcal{H}, \forall \epsilon > 0$, there exists a neighborhood $U$ of $x_0$ such that $g(x) \geq g(x_0) - \epsilon$ for all $x$ in $U$ when $g(x_0) < +\infty$, and $f(x)$ tends to $+\infty$ as $x$ tends towards $x_0$ when $g(x_0) = +\infty$.

Notice that continuity subsumes lower semi-continuity. We also introduce the following metrics for measuring non-convexity from [Natasha].

**Definition 4.** ($\delta$-*Non-Convex Function*) The function $g(x)$ is $\delta$-non-convex if $\forall x, y \in \mathcal{H}$, we have

$$g(y) \geq g(x) + \langle y - x, \nabla g(x)\rangle - \frac{\delta}{2}\|y - x\|^2.$$

Here is our assumption throughout the paper.

**Assumption 1.** Throughout the paper, we make the following assumptions on $f(x)$ and $g(x)$.

- Each $f_i(x)$ is $C^1$, and its gradient $\nabla f_i$ is $L$-Lipschitz continuous.
- Regularizer $g(x)$ is proper, lower semi-continuous and $\delta$-non-convex.
- The minimizer of $F(x)$ is nonempty.

The definition of proximal operator is given below.

**Definition 5.** (*Proximal Operator*) The proximal operator is on $g(x)$ is

$$\text{prox}_{\gamma g}(x) = \arg\min_{y \in \mathbb{R}^d} \left( g(y) + \frac{1}{2\gamma}\|y - x\|^2 \right).$$

Notice that when $\gamma \leq \delta^{-1}$, the function $\hat{g}(y) = g(y) + \frac{1}{2\gamma}\|y - x\|^2$ is proper convex over $\mathcal{H}$, and hence its proximal operator can be effectively computed. We suppose that we are endowed with such an oracle as proximal oracle ($PO$). As in [Natasha, SCSG, ProxSVRG, etc.], we set the norm of the gradient mapping, $\|\mathcal{G}_\gamma(x)\|^2$ as the measurement of our method. Its definition is shown below.

**Definition 6.** (*Gradient Mapping*) The gradient mapping of $x$ for $F(x) = f(x) + g(x)$ is defined as

$$\mathcal{G}_\gamma(x) = \frac{1}{\gamma}\left(x - \text{prox}_{\gamma g}(x - \gamma \nabla f(x))\right).$$

And we call a solution $\hat{x}$ an $\epsilon$-accurate solution if $\mathbb{E}[\|\mathcal{G}_\gamma(\hat{x})\|^2] \leq \epsilon$, where $\hat{x}$ denotes the output of a stochastic algorithm. We also suppose that we are endowed with the following *SFO* oracle.

**Definition 7.** (*Stochastic First-order Oracle*) The *SFO* outputs a stochastic gradient $\nabla f_i(x)$ such that $\mathbb{E}_i(\nabla f_i(x)) = \nabla f(x)$.

We will measure the complexity of our algorithm using the *SFO* complexity.

# 3 SMART+ Algorithm

**Algorithm 1** SMART+

**Require:** choose $\gamma < \delta^{-1}$, $b \leq n$, $q = \frac{b}{n}$, $\tilde{x}_0 = x_0 \in \mathcal{H}$, $\mu_0 = \frac{1}{n}\sum_{i=1}^n (\nabla f_i(\tilde{x}_0))$.

**for** $k = 0, ..., K$ **do**
  sample $I_k \subseteq \{1, ..., n\}$, $p_k \in [0, 1]$;
  **if** $p_k \leq q$ **then**
    $\tilde{x}_{k+1} = x_k$;
    $\mu_{k+1} = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\tilde{x}_{k+1})$;
  **else**
    $\tilde{x}_{k+1} = \tilde{x}_k$;
    $\mu_{k+1} = \mu_k$;
  **end if**
  $x_{k+1} = \text{prox}_{\gamma g}\left(x_k - \gamma\left(\frac{1}{b}\sum_{i \in I_k}(\nabla f_i(x_k) - \nabla f_i(\tilde{x}_{k+1})) + \mu_{k+1}\right)\right)$;
**end for**
**return** uniformly random sample $\hat{x}$ from $\{x_0, ..., x_K\}$.

We have the SMART+ shown above. Note that $I_k$ is a set of cardinality $b$ sampled with replacement from $\{1, ..., n\}$. We make the following standard independence assumption as well.

**Assumption 2.** (*Independence*) The $\sigma$-algebra generated by the history of SMART+, denoted by $\mathcal{F}_k = \sigma((x_0, \tilde{x}_0), ..., (x_k, \tilde{x}_k))$ is independent of the $\sigma$-algebra $\mathcal{I}_k = \sigma((I_k, p_k))$.

Our SMART+ is simple in that it only requires mild changes the SVRG by introducing a randomization mechanism. And it features two significant characteristics compared with previous methods [Natasha, SCSG, etc.].

1. It can handle $\delta$-non-convex lower semi-continuous regularizer.
2. It has better *SFO* complexity in expectation.

# 4 Convergence Results

We first show the results of convergence for SMART+.

**Theorem 1.** (*Convergence Results*) Suppose $\{x_k\}_{k\in\mathbb{N}}$ is generated by SMART+, and all assumptions hold. If we set $\gamma = \min\{\frac{1}{\delta}, \frac{1}{12L}\}$, then the following holds.

- **Object Decrease.** The limit $\lim_{k\to+\infty} F(x_k)$ exists a.s. and for all $k \in \mathbb{N}$, we have

$$\mathbb{E}(F(x_{k+1})|\mathcal{F}_k) \leq F(x_0) - \sum_{t=0}^{k} \left(\frac{\gamma}{12} \left\|\mathcal{G}_\gamma(x_k)\right\|^2\right).$$

- **Limit Points are Stationary.** Suppose that $\{x_k\}_{k\in\mathbb{N}}$ is almost surely bounded. Then $F(\bar{x}_{k+1})$ converges a.s. to a random variable. Moreover, there exists a subset $\tilde{\Omega} \subset \Omega$ such that $P(\tilde{\Omega}) = 1$ and for all $w \in \tilde{\Omega}$, every limiting point of $\{\bar{x}_k\}_{k\in\mathbb{N}}$ is a stationary point.

- **Convergence Rate.** Fix $T \in \mathbb{N}$, and sample $t_0$ uniformly at random from $t_0 \in \{0, ..., T\}$. Then we have

$$\frac{\gamma}{12}\mathbb{E}\left(\left\|\mathcal{G}_\gamma(x_k)\right\|^2\right) \leq \frac{F(x_0) - F(x^*)}{T+1}.$$

  The expected SFO complexity is $\mathbb{E}(N_{SFO}) = \mathcal{O}\left(n + \frac{(L+\delta)b}{\epsilon}\right)$ when $b \geq \sqrt{n}$. Choosing $b = \sqrt{n}$ will lead to complexity of $\mathcal{O}\left(n + \frac{(L+\delta)\sqrt{n}}{\epsilon}\right)$.

Here we provide the convergence analysis for the SMART+. To prove convergence, we hope to bound the norm of the gradient mapping at $x_k$. For all $k \in \mathbb{N}$, we define $\bar{x}_{k+1} \in \mathcal{H}$ as follows.

$$\bar{x}_{k+1} := \text{prox}_{\gamma g}\left(x_k - \frac{\gamma}{n}\sum_{i=1}^{n} \nabla f_i(x_k)\right).$$

Then the gradient mapping is

$$\mathcal{G}_\gamma(x_k) = \frac{1}{\gamma}(x_k - \bar{x}_{k+1}) \in \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_k) + \partial_L g(\bar{x}_{k+1}).$$

In the following subsections we show the proof of this theorem using the supermartingale convergence theorem.

## 4.1 Outline of Proofs

The proof of this theorem is based on the supermartingale convergence theorem shown below.

**Theorem 3.**(*Supermartingale Convergence Theorem*) Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $\Im := \{\mathcal{F}_k\}_{k\in\mathbb{N}}$ be an increasing sequence of sub $\sigma$-algebras of $\mathcal{F}$ such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$. Let $\{X_k\}_{k\in\mathbb{N}}$ and $\{Y_k\}_{k\in\mathbb{N}}$ be sequences of $[\xi,\infty)$-valued and $[0,\infty)$-valued random variables, respectively, such that for all $k \in \mathbb{N}, X_k$ and $Y_k$ are $\mathcal{F}_k$ measurable, and

$$(\forall k \in \mathbb{N}) \qquad \mathbb{E}[X_{k+1}|\mathcal{F}_k] + Y_k \leq X_k.$$

Then $\sum_{k=0}^{\infty} Y_k < \infty$ a.s. and $X_k$ a.s. converges to a $[\xi,\infty)$-valued random variable.

We introduce some definition of variables before moving further.

$$\begin{aligned} R_{\bar{x}}^k =& \|\bar{x}_{k+1} - x_k\|^2 \\ R_x^k =& \|x_{k+1} - x_k\|^2 \\ V_k^i =& \|\nabla f_i(x_k) - \nabla f_i(\tilde{x}_k)\|^2 \\ v_k =& \frac{1}{b}\sum_{i\in I_k}(\nabla f_i(x_k) - \nabla f_i(\tilde{x}_{k+1})) + \mu_{k+1} \end{aligned} \tag{1}$$

Notice that the first and third variable are $\mathcal{F}_k$-measurable, while the other two variables are not. The core of our proof is establishing this supermartingale inequality between the following random variables $X_k$ and $Y_k$.

$$X_k = F(x_k) + \frac{(1-q)}{4nbLq(1+q)}\sum_{i=1}^{n} V_k^i,$$

$$Y_k = \frac{\gamma}{12}\left\|\frac{1}{\gamma}(\bar{x}_{k+1} - x_k)\right\|^2 + \frac{(1-q)}{8nbL}\sum_{i=1}^{n} V_k^i.$$

Without loss of generality, we let $q = \frac{b}{n} \in (0,1)$. Then once the supermartingale inequality is established, we have the following results. There exists a full measure $\tilde{\Omega} \subset \Omega$ such that for all $\omega \in \tilde{\Omega}$, the sequence $\{x_k(\omega)\}_{k\in\mathbb{N}}$ is bounded and

- Because $\sum_{k=0}^{\infty} Y_k < \infty$ a.s., we have $\|\bar{x}_{k+1}(\omega) - x_k(\omega)\|^2 \to 0$ and $\sum_{i=1}^{n} V_k^i \to 0$ as $k \to \infty$.
- Because $X_k \to X_*$ a.s. and $\sum_{i=1}^{n} V_k^i \to 0$, we have $F(x_k(\omega)) \to X_*(\omega)$ as $k \to \infty$.

Then the following lemma shows that the limiting point is also stationary.

**Lemma 1.** Let $\omega \in \tilde{\Omega}$. Suppose that there exits an increasing sequence of indices $\{k_l\}_{l\in\mathbb{N}} \subset \mathbb{N}$ with the property that $\bar{x}_{k_l+1}(\omega) \to \bar{x}(\omega)$, then $F(\bar{x}(\omega)) = X_*(\omega)$. The limit holds that $F(\bar{x}_{k_l+1}(\omega)) \to X_*(\omega) = F(\bar{x}(\omega))$, and there exists $r_{k_l}(\omega) \in \partial_L F(\bar{x}_{k_l+1}(\omega))$ such that $r_{k_l}(\omega) \to 0$ as $l \to \infty$. Therefore we have $0 \in \partial_L F(\bar{x}(\omega))$.

Thus the part 1 and part 2 of Theorem 1 establish once we have proved the supermartingale inequality. And using total expectation, and choose $Y_t$ at random, it's not hard to deduce part 3.

$$\mathbb{E}(Y_t) \leq \frac{1}{(T+1)}\mathbb{E}(X_0 - X_{T+1}) \leq \frac{1}{(T+1)}\mathbb{E}(F(x_0) - F(x^*))$$

## 4.2 Proof of Lemma 1

We first prove Lemma 1, and it's also largely following the convention. Since we have $\|x_{k+1}(\omega) - x_k(\omega)\|^2 \to 0$, thus $\lim_{l\to\infty} x_{k_l}(\omega) = \lim_{l\to\infty} \bar{x}_{k_l+1}(\omega) = \bar{x}(\omega)$, and due to continuity, we have $\lim_{n\to\infty} f(x_{k_l}(\omega)) = \lim_{l\to\infty} f(\bar{x}_{k_l+1}(\omega)) = f(\bar{x}(\omega))$. Since $F(x_{k_l}(\omega)) \to X_*(\omega)$ as $l \to \infty$, thus we have

$$\lim_{l\to\infty} g(x_{k_l}(\omega)) = X_*(\omega) - f(\bar{x}(\omega)).$$

And we have $\|\bar{x}_{k_l+1}(\omega) - x_{k_l}(\omega)\|^2 \to 0$ and $\|x_{k_l}(\omega) - \bar{x}(\omega)\| \to 0$, we hope to prove that $\lim_{n\to\infty} g(x_{k_l}(\omega)) = \lim_{l\to\infty} g(\bar{x}_{k_l+1}(\omega)) = g(\bar{x}(\omega))$. Using that $\{x_{k_l}(\omega)\}_{l\in\mathbb{N}}$ is bounded and the proximal descent lemma, we have

$$g(\bar{x}_{k_l+1}(\omega)) \leq g(x_{k_l}(\omega)) + \left\langle \frac{1}{n}\sum_{i=1}^{n} f_i(x_{k_l}(\omega)), x_{k_l}(\omega) - \bar{x}_{k_l+1}(\omega) \right\rangle - \frac{1}{\gamma}\|\bar{x}_{k_l+1}(\omega) - x_{k_l}(\omega)\|^2.$$

Taking $\lim_{l\to\infty}\sup$ of both sides, and we get

$$\lim_{l\to\infty}\sup g(\bar{x}_{k_l+1}(\omega)) \leq \lim_{l\to\infty} g(x_{k_l}(\omega)).$$

Moreover, for $k_l \geq 1$, we have

$$g(x_{k_l}(\omega)) \leq g(\bar{x}_{k_l+1}(\omega)) + \langle v_{k_l-1}, \bar{x}_{k_l+1}(\omega) - x_{k_l}(\omega)\rangle - \frac{1}{\gamma}\|\bar{x}_{k_l+1}(\omega) - x_{k_l}(\omega)\|^2 \tag{2}$$

Then taking $\lim_{l\to\infty}\inf$ of both sides and we get

$$\lim_{l\to\infty} g(x_{k_l}(\omega)) \leq \lim_{l\to\infty}\inf g(\bar{x}_{k_l+1}(\omega)).$$

This verifies $\lim_{l\to\infty} g(x_{k_l}(\omega)) = \lim_{l\to\infty} g(\bar{x}_{k_l+1}(\omega))$. And using similar arguement, it's not hard to prove that $\lim_{l\to\infty} g(x_{k_l}(\omega)) = g(\bar{x}(\omega))$, since we have $\|x_{k_l}(\omega) - \bar{x}(\omega)\| \to 0$. And finally we have $\lim_{l\to\infty} F(x_{k_l}(\omega)) = \lim_{l\to\infty} F(\bar{x}_{k_l+1}(\omega)) = F(\bar{x}(\omega))$.

## 4.3 Proof of the Supermartingale Inequality

This proof is largely following the convention. Prior to analysis, we state the important lemmas used, and these lemmas will be proved later respectively.

**Lemma 2.** (*Sufficient Decrease*) For all $k \in \mathbb{N}$, we have

$$\begin{aligned}\mathbb{E}_k(F(x_{k+1})) \leq F(x_k) + \left(\frac{L}{2} - \frac{1}{2\gamma}\right)\mathbb{E}_k(R_x^k) + \left(L - \frac{1}{2\gamma}\right)R_{\bar{x}}^k \\ + \mathbb{E}_k\left(\langle x_{k+1} - \bar{x}_{k+1}, \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_k) - v_k\rangle\right).\end{aligned} \tag{3}$$

**Lemma 3.** (*Variance Bound*) For all $k \in \mathbb{N}$, we have

$$\mathbb{E}_k \left( \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_k) - v_k \right\|^2 \right) \leq \frac{1-q}{nb} \sum_{i=1}^{n} V_k^i.$$

**Lemma 4.** (*Dual Recursion*) For all $k \in \mathbb{N}$, and $i \in \{1, ..., n\}$, we have

$$\mathbb{E}_k(V_{k+1}^i) \leq \left( \frac{2}{q} - 1 \right) L^2 \mathbb{E}_k(R_x^k) + \left( 1 - \frac{q(1+q)}{2} \right) V_k^i.$$

The proof of Lemma 2 uses the gradient descent lemma and the proximal descent lemmas. The proof of Lemma 3 and Lemma 4 uses classical probabilistic methods. Based on these lemmas, we now prove the supermartingale inequality as follows.

First, we have the following bound.

$$\begin{aligned} &\mathbb{E}_k \left( \langle x_{k+1} - \bar{x}_{k+1}, \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_k) - v_k \rangle \right) \\ &\leq 2L \mathbb{E}_k \left( \|x_{k+1} - \bar{x}_{k+1}\|^2 \right) + \frac{1}{8L} \mathbb{E}_k \left( \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_k) - v_k \right\|^2 \right) \\ &\leq 4L \mathbb{E}_k(R_x^k) + 4L R_{\bar{x}}^k + \frac{1-q}{8nbL} \sum_{i=1}^{n} V_k^i \end{aligned} \tag{4}$$

We used Young's inequality in the proof. Insert this inequality into the original inequality and use $\gamma \leq \frac{1}{12L}$, we have

$$\mathbb{E}_k(F(x_{k+1})) \leq F(x_k) - L R_{\bar{x}}^k - \frac{3L}{2} \mathbb{E}_k \left( R_x^k \right) + \frac{1-q}{8nbL} \sum_{i=1}^{n} V_k^i.$$

And by definition of $X_k$, we have

$$\begin{aligned} \mathbb{E}_k(X_{k+1}) \leq X_k - L R_{\bar{x}}^k - \frac{3L}{2} \mathbb{E}_k \left( R_x^k \right) \\ + \frac{1-q}{8nbL} \sum_{i=1}^{n} V_k^i + \frac{(1-q)}{4nbLq(1+q)} \sum_{i=1}^{n} (\mathbb{E}_k(V_{k+1}^i) - V_k^i) \end{aligned} \tag{5}$$

We use the Lemma 4 to get

$$\mathbb{E}_k(X_{k+1}) \leq X_k - L R_{\bar{x}}^k - \left( \frac{3L}{2} - \frac{n(n-b)(2n-b)L}{2b^3(n+b)} \right) \mathbb{E}_k \left( R_x^k \right) - \frac{1-q}{8nbL} \sum_{i=1}^{n} V_k^i \tag{6}$$

And when $b \geq n^{\frac{2}{3}}$, we have

$$\frac{3L}{2} - \frac{n(n-b)(2n-b)L}{2b^3(n+b)} \geq \frac{3L}{2} - L > 0.$$

Thus we have

$$\mathbb{E}_k(X_{k+1}) \leq X_k - \frac{\gamma}{12}\left\|\mathcal{G}_\gamma(x_k)\right\|^2 - \frac{1-q}{8nbL}\sum_{i=1}^{n} V_k^i = X_k - Y_k. \tag{7}$$

This establishes the supermartingale inequality. Micro-scoping over this inequality and taking expectations over all $\mathcal{I}_k$ gives us

$$\mathbb{E}(X_{T+1}) \leq X_0 - \frac{\gamma}{12}\sum_{k=0}^{T}\mathbb{E}\left(\left\|\mathcal{G}_\gamma(x_k)\right\|^2\right)$$

Since $X_0 = F(x_0)$, and $\mathbb{E}(X_{T+1}) \geq \mathbb{E}(F(x_{T+1})) \geq F(x^*)$, if we choose $\hat{x}$ from $\{x_0, ..., x_T\}$ uniformly at random, then we have

$$\mathbb{E}\left(\left\|\mathcal{G}_\gamma(\hat{x})\right\|^2\right) \leq \frac{12(F(x_0) - F(x^*))}{\gamma(T+1)} \leq \epsilon.$$

This leads to

$$T \geq \frac{12(F(x_0) - F(x^*))}{\gamma\epsilon}.$$

And the expected *SFO* is

$$n + 3bT = n + \frac{36b(F(x_0) - F(x^*))}{\gamma\epsilon} = \mathcal{O}\left(n + \frac{(L+\delta)b}{\epsilon}\right).$$

Letting $b = n^{\frac{2}{3}}$ leads to *SFO* complexity of $\mathcal{O}\left(n + \frac{(L+\delta)n^{\frac{2}{3}}}{\epsilon}\right)$.

## 4.4 Proof of Lemma 2-4

### 4.4.1 Proof of Lem2.

Using descent lemma, we have

$$f_i(\bar{x}_{k+1}) \leq f_i(x_k) + \langle \bar{x}_{k+1} - x_k, \nabla f_i(x_k)\rangle + \frac{L}{2}R_{\bar{x}}^k.$$

For proximal gradient, we have

$$g(\bar{x}_{k+1}) \leq g(x_k) + \langle x_k - \bar{x}_{k+1}, \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_k)\rangle - \frac{1}{\gamma}R_{\bar{x}}^k.$$

Adding the inequalities, then we get

$$F(\bar{x}_{k+1}) \leq F(x_k) + \left(\frac{L}{2} - \frac{1}{\gamma}\right)R_{\bar{x}}^k.$$

Also we have for $x_{k+1}$, the descent lemma yields

$$f_i(x_{k+1}) \leq f_i(x_k) + \langle x_{k+1} - x_k, \nabla f_i(x_k)\rangle + \frac{L}{2}R_x^k,$$

$$f_i(x_k) \leq f_i(\bar{x}_{k+1}) + \langle x_k - \bar{x}_{k+1}, \nabla f_i(x_k)\rangle + \frac{L}{2} R_{\bar{x}}^k.$$

And the proximal optimality yields

$$g(x_{k+1}) + \langle v_k, x_{k+1}\rangle + \frac{1}{2\gamma} R_x^k \leq g(\bar{x}_{k+1}) + \langle v_k, \bar{x}_{k+1}\rangle + \frac{1}{2\gamma} R_{\bar{x}}^k.$$

Adding these three inequalities and we have

$$\begin{aligned} F(x_{k+1}) \leq F(\bar{x}_{k+1}) + \left(\frac{L}{2} - \frac{1}{2\gamma}\right) R_x^k + \left(\frac{L}{2} + \frac{1}{2\gamma}\right) R_{\bar{x}}^k \\ + \langle x_{k+1} - \bar{x}_{k+1}, \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_k) - v_k\rangle. \end{aligned} \tag{8}$$

Finally adding the previous two inequalities, we have

$$\begin{aligned} \mathbb{E}_k(F(x_{k+1})) \leq F(x_k) + \left(\frac{L}{2} - \frac{1}{2\gamma}\right) \mathbb{E}_k(R_x^k) + \left(L - \frac{1}{2\gamma}\right) R_{\bar{x}}^k \\ + \mathbb{E}_k\left(\langle x_{k+1} - \bar{x}_{k+1}, \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_k) - v_k\rangle\right). \end{aligned} \tag{9}$$

#### 4.4.2 Proof of Lem3.

Define $\epsilon_k^i = \nabla f_i(x_k) - \nabla f_i(\tilde{x}_k)$, then we have

$$\begin{aligned}
&\mathbb{E}_k\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_k) - v_k\right\|^2\right) \\
&= (1-q)\mathbb{E}_k\left(\left\|\frac{1}{b}\sum_{j\in I_k}\left[(\nabla f_i(x_k) - \nabla f_i(\tilde{x}_k)) - \frac{1}{n}\sum_{i=1}^{n}(\nabla f_i(x_k) - \nabla f_i(\tilde{x}_k))\right]\right\|^2\right) \\
&= (1-q)\mathbb{E}_k\left(\left\|\frac{1}{b}\sum_{j\in I_k}\left[\epsilon_k^j - \mathbb{E}_{k,l\sim Unif[n]}[\epsilon_k^l]\right]\right\|^2\right) \\
&= \frac{1-q}{b^2}\mathbb{E}_k\left(\sum_{j\in I_k}\left\|\left[\epsilon_k^j - \mathbb{E}_{k,l\sim Unif[n]}[\epsilon_k^l]\right]\right\|^2\right) \\
&= \frac{1-q}{b^2}\mathbb{E}_k\left(\sum_{j\in I_k}\left[\|\epsilon_k^j\|^2 - \|\mathbb{E}_{k,l\sim Unif[n]}[\epsilon_k^l]\|^2\right]\right) \\
&\leq \frac{1-q}{nb}\sum_{i=1}^{n}\|\epsilon_k^i\|^2 \\
&= \frac{1-q}{nb}\sum_{i=1}^{n}V_k^i.
\end{aligned} \tag{10}$$

#### 4.4.3 Proof of Lem4.

Since we have

$$\begin{aligned}
\mathbb{E}_k(V_{k+1}^i) &= \mathbb{E}_k(\|\nabla f_i(x_{k+1}) - \nabla f_i(\tilde{x}_{k+1})\|^2) \\
&= q\mathbb{E}_k(\|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\|^2) + (1-q)\mathbb{E}_k(\|\nabla f_i(x_{k+1}) - \nabla f_i(\tilde{x}_k)\|^2) \\
&\leq \left(q + (1-q)\left(1+\frac{2}{q}\right)\right)\mathbb{E}_k(\|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\|^2) + (1-q)\left(1+\frac{q}{2}\right)V_k^i \\
&\leq \left(\frac{2}{q} - 1\right)L^2\mathbb{E}_k(R_x^k) + \left(1 - \frac{q(1+q)}{2}\right)V_k^i
\end{aligned} \tag{11}$$

Here we used inequality $\|a+b\|^2 \leq (1+q^{-1})\|a\|^2 + (1+q)\|b\|^2$.

# 5 Online Variant

In this section, we consider a online variant of the SMART+, in fact it only have some subtle change in setting the batch size. The algorithm is shown below.

**Algorithm 2** SMART+ Online

**Require:** choose $\gamma < \delta^{-1}$, $B = \frac{100\sigma^2}{\epsilon}$, $b \leq B$, $q = \frac{b}{B}$, $\tilde{x}_0 = x_0 \in \mathcal{H}$, $\mu_0 = \frac{1}{B}\sum_{i\in B_0}(\nabla f_i(\tilde{x}_0))$.

**for** $k = 0, ..., K$ **do**
  sample $I_k \subseteq \{1, ..., n\}$, $p_k \in [0, 1]$, $B_k \subseteq \{1, ..., n\}$;
  **if** $p_k \leq q$ **then**
    $\tilde{x}_{k+1} = x_k$;
    $\mu_{k+1} = \frac{1}{B}\sum_{i\in B_k} \nabla f_i(\tilde{x}_{k+1})$;
  **else**
    $\tilde{x}_{k+1} = \tilde{x}_k$;
    $\mu_{k+1} = \mu_k$;
  **end if**
  $x_{k+1} = \text{prox}_{\gamma g}\left(x_k - \gamma\left(\frac{1}{b}\sum_{i\in I_k}(\nabla f_i(x_k) - \nabla f_i(\tilde{x}_{k+1})) + \mu_{k+1}\right)\right)$;
**end for**
**return** uniformly random sample $\hat{x}$ from $\{x_0, ..., x_K\}$.

One thing to notice is that, $I_k$ is sampled with replacement, while $B_k$ is a set of cardinality $B$ sampled without replacement from $\{1, ..., n\}$. We make the standard assumption on the variance of the stochastic derivate.

**Assumption 3.** (*Stochastic Derivative Variance*) $\forall x \in \mathcal{H}$, $\mathbb{E}_i(\|\nabla f_i(x) - \nabla f(x)\|^2) \leq \sigma^2$.

We also assume that the $\sigma$-algebra generated by SMART+ online, $\mathcal{F}_k$, is independent with $\mathcal{I}'_k = \{(I_k, p_k, B_k)\}$. We also make the following standard assumption for online settings.

**Assumption 4.** (*Online Assumption*) $\epsilon^{-1} = o(n)$, or equivalently, $n \gg B = \mathcal{O}(\frac{1}{\epsilon})$.

Then we have the following theorem for the performance of SMART+ online.

**Theorem 2** (*Convergence for SMART+ Online*) For SMART+ online, if we set the parameters as $\gamma = \min\{\frac{1}{12L}, \frac{1}{\delta}\}$, $B = \frac{125\sigma^2}{\epsilon}$, $b \geq B^{\frac{2}{3}}$ then we have expected *SFO* complexity of $\mathcal{O}\left(\frac{(L+\delta)b}{\epsilon}\right)$ for achieving an $\epsilon$-optimal solution. And if $b = \frac{25\sigma^{\frac{4}{3}}}{\epsilon^{\frac{2}{3}}}$, we have expected SFO complexity of

$$\mathcal{O}\left(\frac{(L+\delta)\sigma^{\frac{4}{3}}}{\epsilon^{\frac{5}{3}}}\right).$$

## 5.1 Proof Outline

Most auxiliary variables are defined in agreement with the offline case, except for the $\mu_{k+1}$.

$$\mu_{k+1} = \frac{1}{B}\sum_{j\in B_k} \nabla f_j(x_k)$$

Before progress to the proofs, we first observe that, Lemma 2 and Lemma 4 also establish in the online case. However, to obtain the convergence result of SMART+ online, we need a modified

version of Lemma 3 as below.

**Lemma 5.**(*Modified Variance Bound*) For all $k \in \mathbb{N}$, we have

$$\mathbb{E}_k\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_k) - v_k\right\|^2\right) \leq \frac{1-q}{nb}\sum_{i=1}^{n}V_k^i + \frac{\sigma^2}{B}.$$

Here we prove our main theorem. Since we have

$$\begin{aligned}
&\mathbb{E}_k\left(\langle x_{k+1} - \bar{x}_{k+1}, \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_k) - v_k\rangle\right) \\
&\leq 2L\mathbb{E}_k\left(\|x_{k+1} - \bar{x}_{k+1}\|^2\right) + \frac{1}{8L}\mathbb{E}_k\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_k) - v_k\right\|^2\right) \\
&\leq 4L\mathbb{E}_k(R_x^k) + 4LR_{\bar{x}}^k + \frac{1-q}{8nbL}\sum_{i=1}^{n}V_k^i + \frac{\sigma^2}{8LB}
\end{aligned} \tag{12}$$

We substitute this inequality into Lemma 2 and get

$$\mathbb{E}_k(F(x_{k+1})) \leq F(x_k) - LR_{\bar{x}}^k - \frac{3L}{2}\mathbb{E}_k\left(R_x^k\right) + \frac{1-q}{8nbL}\sum_{i=1}^{n}V_k^i + \frac{\sigma^2}{8LB} \tag{13}$$

And by definition of $X_k$, we have

$$\begin{aligned}
\mathbb{E}_k(X_{k+1}) \leq X_k - LR_{\bar{x}}^k - \frac{3L}{2}\mathbb{E}_k\left(R_x^k\right) + \frac{1-q}{8nbL}\sum_{i=1}^{n}V_k^i \\
+ \frac{\sigma^2}{8LB} + \frac{1-q}{2nbLq(1+q)}\sum_{i=1}^{n}(\mathbb{E}_k(V_{k+1}^i) - V_k^i)
\end{aligned} \tag{14}$$

We use Lemma 4 to get

$$\mathbb{E}_k(X_{k+1}) \leq X_k - LR_{\bar{x}}^k - \left(\frac{3L}{2} - \frac{B^2(B-b)(2B-b)L}{2nb^3(B+b)}\right)\mathbb{E}_k\left(R_x^k\right) - \frac{1-q}{8nbL}\sum_{i=1}^{n}V_k^i + \frac{\sigma^2}{8BL}.$$

And when $b \geq B^{\frac{2}{3}}$, we have

$$\frac{3L}{2} - \frac{B^2(B-b)(2B-b)L}{2nb^3(B+b)} \geq \frac{3L}{2} - L > 0.$$

Notice we used the presumption that $n \gg B = \mathcal{O}(\frac{1}{\epsilon})$. Thus we have

$$\mathbb{E}_k(X_{k+1}) \leq X_k - \frac{\gamma}{12}\left\|\mathcal{G}_\gamma(x_k)\right\|^2 - \frac{1-q}{8nbL}\sum_{i=1}^{n}V_k^i + \frac{\sigma^2}{8BL}. \tag{15}$$

By microscoping and take expectation of all time we get

$$\mathbb{E}(X_{T+1}) \leq \mathbb{E}(X_0) - \sum_{k=0}^{T} \frac{\gamma}{12} \left\|\mathcal{G}_\gamma(x_k)\right\|^2 + \frac{\sigma^2(T+1)}{8BL}.$$

And then we have

$$\mathbb{E}(F(x_T)) \leq F(x_0) - \sum_{k=0}^{T} \frac{\gamma}{12} \left\|\mathcal{G}_\gamma(x_k)\right\|^2 + \frac{\sigma^2(T+2)}{8BL}.$$

Taking random $x_k$ over $k$, we have

$$\mathbb{E}(\left\|\mathcal{G}_\gamma(x_k)\right\|^2) \leq \frac{12(F(x_0) - F(x^*))}{\gamma(T+1)} + \frac{18\sigma^2(T+2)}{B(T+1)}.$$

Since we set $B = \frac{125\sigma^2}{\epsilon} = \mathcal{O}(\frac{1}{\epsilon})$, $T = \frac{24(F(x_0)-F(x^*))}{\gamma\epsilon}$, then we have

$$\mathbb{E}(\left\|\mathcal{G}_\gamma(x_k)\right\|^2) \leq \frac{\epsilon}{2} + \frac{36\sigma^2}{B} \leq \epsilon.$$

And the expected number of *SFO* is $3(T+2)b = \mathcal{O}(\frac{(L+\delta)b}{\epsilon})$. Let $b = \frac{25\sigma^{\frac{4}{3}}}{\epsilon^{\frac{2}{3}}}$, and we get *SFO* complexity of

$$\mathcal{O}\left(\frac{(L+\delta)\sigma^{\frac{4}{3}}}{\epsilon^{\frac{5}{3}}}\right).$$

### 5.2 Proof for lemma 5.

Define $\epsilon_k^i = \nabla f_i(x_k) - \nabla f_i(\tilde{x}_k)$, then we have

$$
\begin{aligned}
&\mathbb{E}_k \left( \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_k) - v_k \right\|^2 \right) \\
&= (1-q)\mathbb{E}_k \left( \left\| \frac{1}{b} \sum_{j \in I_k} \left[ (\nabla f_i(x_k) - \nabla f_i(\tilde{x}_k)) - (\nabla f(x_k) - \nabla f(\tilde{x}_k)) \right] + \frac{1}{B} \sum_{j \in \mathcal{B}_k} (\nabla f_j(\tilde{x}_k) - \nabla f(\tilde{x}_k)) \right\|^2 \right) \\
&+ q \cdot \mathbb{E}_k \left( \left\| \frac{1}{B} \sum_{j \in \mathcal{B}_k} (\nabla f_j(x_k) - \nabla f(x_k)) \right\|^2 \right) \\
&= (1-q)\mathbb{E}_k \left( \left\| \frac{1}{b} \sum_{j \in I_k} \left[ \epsilon_k^j - \mathbb{E}_{k, l \sim Unif[n]} [\epsilon_k^l] \right] \right\|^2 \right) + (1-q) \cdot \mathbb{E}_k \left( \left\| \frac{1}{B} \sum_{j \in \mathcal{B}_k} (\nabla f_j(\tilde{x}_k) - \nabla f(\tilde{x}_k)) \right\|^2 \right) \\
&+ q \cdot \mathbb{E}_k \left( \left\| \frac{1}{B} \sum_{j \in \mathcal{B}_k} (\nabla f_j(x_k) - \nabla f(x_k)) \right\|^2 \right) \\
&\leq \frac{1-q}{nb} \sum_{i=1}^{n} V_k^i + \frac{\sigma^2}{B}.
\end{aligned}
\tag{16}
$$

# 6 Conclusion

In this paper, we considered the fully non-convex composite optimization problem. And we designed SMART+, which introduced a randomized mechanism into ProxSVRG. Our algorithm only requires mild changes over the ProxSVRG method, and our analysis yields a competitive complexity upper bound for the non-convex composite optimization. The online variant of SMART+ also achieves competitive complexity compared with previous online methods.

# 7 References


1. Zhize Li, Jian Li. A simple proximal stochastic gradient method for non-smooth non-convex optimization.

2. Blake Woodworth, Nathan Srebro. Tight complexity bounds for optimizing composite objectives.

3. Jerome Bolte, Shoham Sabach, Marc Teboulle. Proximal alternating linearized minimization for non-convex and non-smooth problems.

4. Hedy Attouch, Jerome Bolte, Benar Fux Svaiter. Convergence of decent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods.

5. Damek Davis. The asynchronous PALM algorithm for non-smooth non-convex problems.

6. Aleksandr Aravkin, Damek Davis. A SMART stochastic algorithm for non-convex optimization with applications to robust machine learning.

7. Zeyuan Allen-Zhu. Katyusha: the first direct acceleration of stochastic gradient methods.

8. Zeyuan Allen-Zhu. Natasha: faster non-convex stochastic optimization via strongly non-convex parameter.

9. Sashank J. Reddi, Suvrit Sra, Barnabas Poczos, Alex Smola. Fast stochastic methods for non-smooth non-convex optimiziation.

10. Rie Johnson, Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction.

11. Lihua Lei, Cheng Ju, Jianbo Chen, Micheal I. Jordan. Non-convex finite-sum optimization via SCSG methods.