SVRG-SO: SVRG for Stochastic Optimization[*]

Sourav Sahoo[1]

[1]Department of Electrical Engineering, Indian Institute of Technology Madras
sourav.sahoo@smail.iitm.ac.in

Abstract

Stochastic Variance Reduced Gradient (SVRG) belongs to a family of gradient aggregation algorithms widely used in large-scale machine learning applications. SVRG achieved linear convergence for strongly convex finite-sum optimization problems. Although similar results were also obtained by previous algorithms such as stochastic dual coordinate ascent (SDCA) and stochastic average gradient (SAG), SVRG is more intuitive and simpler to analyse. Furthermore, SVRG also works well in practice for non-strongly convex or non-convex finite-sum problems. Owing to such robustness and simplicity of the algorithm, we adapt it for stochastic optimization problems in this work. In this pursuit, we present SVRG for Stochastic Optimization (**SVRG-SO**) and recover optimal convergence rates for strongly convex and smooth objective functions.

1 INTRODUCTION

We consider a stochastic optimization (SO) problem of the form

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f(x) := \mathbb{E}_\xi[F(x,\xi)] \tag{SO}$$

where ξ is a random vector whose probability distribution is supported on Ξ and $F : \mathbb{R}^d \times \Xi \to \mathbb{R}$ is a real-valued function. When the training data is finite, this problem can be transformed into the finite sum optimization problem which is of the form:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \tag{FS}$$

where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is a real-valued function. In the context of statistical learning theory and machine learning, the stochastic optimization problem and finite sum problem are commonly known as the expected risk minimization and empirical risk minimization problem respectively.

The finite sum problem has received considerable attention in the past decade due to the exponential growth of large-scale machine learning applications. In this regard, several incremental methods [DD$^+$14, DBLJ14] and accelerated methods [ZDS$^+$19, SSZ14] have been proposed. However, for better generalization of the proposed model, it is essential to minimize the *expected risk* over the true (possibly unknown) distribution of data (which is equivalent to minimize (SO)) instead of minimizing *empirical risk* (which is equivalent to minimize (FS)) [BB07]. Hence, it is essential to bridge the gap between them.

[*]Final project for CS6515 Stochastic Optimization, Indian Institute of Technology Madras, Spring 2022.

Stochastic gradient descent (SGD) has been the de facto optimization algorithm for most machine learning tasks. However, due to the variance in random sampling, it can provide only $\mathcal{O}\left(1/k\right)$ convergence. To alleviate this issue, several variance reduction methods such as SVRG [JZ13], SAGA [DBLJ14] and SAG [SLRB17] have been proposed. In this work, we adapt SVRG for stochastic optimization and present the **SVRG-SO** algorithm.

1.1 Notation and Preliminaries

Gradient Noise. When f has a finite-sum structure, several papers analysing SGD rely on the assumption that $\mathbb{E}\left[\left\|\nabla f_i(x)\right\|^2\right] \leq c < \infty$ [HK14, RRWN11]. However, it is easy to see that for strongly convex unconstrained optimization, such an assumption is not valid. Instead a weaker assumption, $\mathbb{E}\left[\left\|\nabla f_i(x)\right\|^2\right] \leq c_1 + c_2\mathbb{E}\left[\left\|\nabla f(x)\right\|^2\right]$ is adopted. In the case of finite sum problem, when $c_1 = 0$, it is known as Strong Growth Condition (SGC). Under SGC, stochastic gradient method with fixed step size achieves linear convergence [SR13]. Cevher and Vu [CV19] later adapted the condition for (SO) and provided the necessary and sufficient conditions for linear convergence of SGD, i.e., $\mathbb{E}\left[\left\|\nabla F(x,\xi)\right\|^2\right] \leq c_1 + c_2\mathbb{E}\left[\left\|\nabla f(x)\right\|^2\right]$. When $c_1 > 0$, it is known as Weak Growth Condition (WGC) of f, and if $c_1 = 0$, it is called Growth Condition (GC) of f.

$\left\|\cdot\right\|$ denotes the Euclidean norm, unless stated otherwise. $x \lesssim y$ implies there exists some uniform constant C, such that $x \leq Cy$.

Table 1: Convergence guarantee of the proposed algorithm in terms of number of iterations k. For simplicity, we suppress the dependence on strong convexity, Lipschitz smoothness parameters etc. here. The exact dependence in provided in Theorem 1 and Theorem 2. f: objective function, η_t: step size for iteration t.

	Growth Condition of f	Weak Growth Condition of f
Fixed Step Size	$\mathcal{O}\left(\rho^k\right), 0 < \rho < 1$	$\mathcal{O}\left(\rho^k\right) + \mathcal{O}\left(1\right), 0 < \rho < 1$
Diminishing Step Size $\left\{\eta_t = \mathcal{O}\left(\frac{1}{t}\right)\right\}$	$\mathcal{O}\left(\frac{1}{k^c}\right), c > 1$	$\mathcal{O}\left(\frac{1}{k}\right)$

2 MAIN RESULTS

In this section, we present the **SVRG-SO** algorithm which depends on the following assumptions:

Assumption 1 (L-smoothness)**.** *We assume that f is L-smooth, i.e.,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\left\|y - x\right\|^2, \forall x, y \tag{1}$$

Assumption 2 (μ-strong Convexity)**.** *We assume that f is μ-strongly convex, i.e.,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\left\|y - x\right\|^2, \forall x, y \tag{2}$$

Define $\mathcal{F}_k^j = \{X_{k,1}, \xi_1', X_{k,2}, \xi_2', \ldots, \xi_{j-1}', X_{k,j}\}$ as the σ-algebra generated by the observations of the first j inner loop iterations in Algorithm 1. We assume that both the $\{\xi_j\}_{j=1}^{n_k}$ and $\{\xi_j'\}_{j=1}^{m}$ are i.i.d. random variables. We further assume that we have access to a stochastic first-order oracle which satisfies the following conditions:

Assumption 3 (Unbiased Gradients)**.** *The stochastic gradients computed at X is an unbiased estimator of the gradient at that point, i.e., $\mathbb{E}[\nabla F(X_k, \xi)] = \nabla f(X_k)$ and $\mathbb{E}[\nabla F(X_{k,j}, \xi_j')|\mathcal{F}_k^j] = \nabla f(X_{k,j})$.*

Assumption 4 (Affine Variance)**.** *We assume that* $\mathbb{E}\left[\|\nabla F(X_k,\xi)-\nabla f(X_k)\|^2\right] \leq \sigma_0^2+\sigma_1^2\|\nabla f(X_k)\|^2$ *and* $\mathbb{E}\left[\left\|\nabla F(X_{k,j},\xi_j')-\nabla f(X_{k,j})\right\|^2|\mathcal{F}_k^j\right] \leq \sigma_0^2+\sigma_1^2\|\nabla f(X_{k,j})\|^2$, $\forall j\in[m]$ *for some* $\sigma_0\geq 0,\sigma_1\geq 0$*. Hence,*

$$\mathbb{E}\left[\|\nabla F(X_k,\xi)\|^2\right] \leq \sigma_0^2+(1+\sigma_1^2)\|\nabla f(X_k)\|^2 \tag{3}$$

$$\mathbb{E}\left[\|\nabla F(X_{k,j},\xi_j')\|^2|\mathcal{F}_k^j\right] \leq \sigma_0^2+(1+\sigma_1^2)\|\nabla f(X_{k,j})\|^2 \tag{4}$$

Remark 1. *We emphasize that we do not impose any restrictions on* $F(X,\xi)$ *except that it is differentiable. The strong convexity and smoothness conditions are only for the objective function* $f(x)$*.*

Remark 2. *It is essential to consider affine variance bounds for stochastic gradients to correctly model the problem in machine learning tasks with feature noise or in situations where the model parameters are multiplicatively perturbed by noise, such as a deep neural network [FTC+22].*

Algorithm 1 SVRG-SO

```
Choose $\eta_k > 0$, $m \in \mathbb{N}$ and initialize $X_1 \in \mathbb{R}^d$.
for $k = 1, 2, \ldots$ do
    $G(X_k) = \frac{1}{n_k}\sum_{j=1}^{n_k} \nabla F(X_k, \xi_j)$
    $X_{k,1} = X_k$.
    for $j = 1, 2, \ldots, m$ do
        $g_j = \nabla F(X_{k,j}, \xi_j') - (\nabla F(X_k, \xi_j') - G(X_k))$
        $X_{k,j+1} = X_{k,j} - \eta_k g_j$.
    end for
    $X_{k+1} = X_{k,m+1}$
end for
```

2.1 Convergence Analysis

We first state an important lemma which would be essential for the main results. All the proofs in this section are deferred to the appendix.

Lemma 1. *Suppose Assumption 1, 2, 3 and 4 hold true. Let* x^* *be the optimal point. Furthermore, assume that the step size* η_k *satisfies*

$$0 < \eta_k \leq \frac{1}{4L(1+\sigma_1^2)}$$

Denote the expected optimality gap at instance t following Algorithm 1 as Δ_t*, i.e.,* $\Delta_t = \mathbb{E}[f(X_t)] - f(x^*)$*. Then,*

$$\Delta_{k,j+1} \leq \left(1-2\mu\eta_k(1-2\eta_k L(1+\sigma_1^2)\right)\Delta_{k,j} + L\eta_k^2\left(\left(4+\frac{1}{n_k}\right)\sigma_0^2+\frac{2L^2}{\mu}\left(3+2\sigma_1^2+\frac{\sigma_1^2}{n_k}\right)\Delta_k\right) \tag{5}$$

Theorem 1 (strongly convex, fixed step size)**.** *Suppose Assumption 1, 2, 3 and 4 hold true. Let* x^* *be the optimal point. Furthermore, assume that the fixed step size* η *and* $m \in \mathbb{N}$ *satisfies*

$$0 < \eta \leq \frac{\mu^2}{12L^3(1+\sigma_1^2)} \qquad \textit{and} \qquad m \geq \left\lceil \frac{3}{2\mu\eta} \right\rceil$$

Then, for,

$$\rho = \frac{1}{1-2\eta L(1+\sigma_1^2)}\left(\frac{1}{1+2m\mu\eta} + \frac{3L^3\eta(1+\sigma_1^2)}{\mu^2}\right) \in (0,1) \tag{6}$$

we have the two results:

(i) for growth condition of f, i.e., ($\sigma_0 = 0$)

$$\mathbb{E}\left[f(X_{k+1}) - f(x^*)\right] \leq \rho^k \mathbb{E}\left[f(X_1) - f(x^*)\right] \tag{7}$$

Hence, for any given $\varepsilon > 0$,

$$k \geq \frac{1}{1-\rho}\log\left(\frac{\mathbb{E}\left[f(X_1) - f(x^*)\right]}{\varepsilon}\right) \tag{8}$$

implies $\mathbb{E}\left[f(X_{k+1}) - f(x^*)\right] \leq \varepsilon$.

(ii) for weak growth condition of f, i.e., ($\sigma_0 > 0$)

$$\mathbb{E}\left[f(X_{k+1}) - f(x^*)\right] \leq \rho^k \mathbb{E}\left[f(X_1) - f(x^*)\right] + \frac{5L\eta\sigma_0^2}{\mu(1-\rho)} \tag{9}$$

Hence, for any given $\varepsilon > 0$, choosing step size

$$\eta \leq \min\left\{\frac{\mu^2}{12L^3(1+\sigma_1^2)}, \frac{\varepsilon\mu}{30L\sigma_0^2}\right\} \tag{10}$$

and

$$k \geq \frac{1}{1-\rho}\log\left(\frac{2\mathbb{E}\left[f(X_1) - f(x^*)\right]}{\varepsilon}\right) \tag{11}$$

implies $\mathbb{E}\left[f(X_{k+1}) - f(x^*)\right] \leq \varepsilon$.

Theorem 2 (strongly convex, diminishing step size)**.** *Suppose Assumption 1, 2, 3 and 4 hold true. Let x^* be the optimal point. Furthermore, assume that the step size η_k and $m \in \mathbb{N}$ is given as:*

$$\eta_k = \frac{\theta}{\lambda + k} \quad \textit{where} \quad \theta > \frac{5}{2m\mu} \quad \textit{and} \quad \lambda > 0 \quad \textit{such that} \quad \eta_1 \leq \frac{\mu^2}{12L^3(1+\sigma_1^2)} \quad \textit{and} \quad m \leq \left\lfloor \frac{2}{5\mu\eta_1} \right\rfloor \tag{12}$$

Then, we have the two results:

(i) for growth condition of f, i.e., ($\sigma_0 = 0$)

$$\mathbb{E}\left[f(X_{k+1}) - f(x^*)\right] \leq \left(\frac{\lambda+1}{\lambda+k}\right)^c \mathbb{E}\left[f(X_1) - f(x^*)\right] \tag{13}$$

where $c = \frac{2}{5}m\mu\theta > 1$. Hence, for any given $\varepsilon > 0$,

$$k \geq \left(\frac{\mathbb{E}\left[f(X_1) - f(x^*)\right]}{\varepsilon}\right)^{1/c}(\lambda + 1) - \lambda \tag{14}$$

implies $\mathbb{E}\left[f(X_{k+1}) - f(x^*)\right] \leq \varepsilon$.

(ii) for weak growth condition of f, i.e., ($\sigma_0 > 0$)

$$\mathbb{E}\left[f(X_{k+1}) - f(x^*)\right] \leq \frac{\nu}{\lambda + k} \tag{15}$$

where

$$\nu = \max\left\{(\lambda+1)\mathbb{E}\left[f(X_1) - f(x^*)\right], \frac{30mL\sigma_0^2\theta^2}{2m\mu\theta - 5}\right\}$$

Hence, for any given $\varepsilon > 0$,

$$k \geq \frac{\nu}{\varepsilon} - \lambda \tag{16}$$

implies $\mathbb{E}\left[f(X_{k+1}) - f(x^*)\right] \leq \varepsilon$.

Discussion. From Theorem 1 and Theorem 2, it is clear that the rate of convergence largely depends on the guarantees ensured about gradient noise. Under GC of f, choosing fixed step size allows to obtain an ε-optimal solution in $\mathcal{O}\left(\log(1/\varepsilon)\right)$ outer loop iterations. Under WGC of f, although the fixed step variant ensures linear convergence to the vicinity of the solution, it does not converge to it. The diminishing step size variant converges to the optimal solution but at a "slow" rate. So, by combining the best of both worlds, it is possible to obtain convergence much faster in practice in the following way: a) obtain an ε-optimal solution using fixed step size and then use it as the initial point, use diminishing step size to reach the optimal value.

Furthermore, we observe that the dependence of the convergence rate on the number of data points, n_k, used to estimate $G(X_K)$ (line 3 of Algorithm 1) is negligible. A tighter analysis (see Appendix C) might obtain a convergence rate depending on n_k. However, it can be verified that the new convergence rate is better than the present rate by, at best, a small constant factor.

3 CONCLUSION

In this work, we proposed and analysed **SVRG-SO** algorithm. For strongly convex, L-smooth functions, under GC, we obtained $\mathcal{O}\left(\rho^k\right)$ convergence rate for the fixed-step variant and $\mathcal{O}\left(k^{-c}\right), c > 1$ rate for diminishing step sizes. Furthermore, under WGC, we get $\mathcal{O}\left(1/k\right)$ rate for diminishing step sizes. However, for fixed step size, we observe that although the algorithm converges linearly to the proximity of the optimal value, there is an asymptotic residual term. In the future, we intend to extend this work for non-strongly convex functions and even non-smooth objective functions.

ACKNOWLEDGEMENTS

The author would like to thank Raghu Pasupathy for his meaningful feedback on the initial draft of this work.

References

[BB07] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. *Advances in neural information processing systems*, 20, 2007. 1

[BCN18] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018. 14

[CV19] Volkan Cevher and Bang Công Vũ. On the linear convergence of the stochastic gradient method with constant step-size. *Optimization Letters*, 13(5):1177–1187, 2019. 2

[DBLJ14] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014. 1, 2

[DD$^+$14] Aaron Defazio, Justin Domke, et al. Finito: A faster, permutable incremental gradient method for big data problems. In *International Conference on Machine Learning*, pages 1125–1133. PMLR, 2014. 1

[FTC$^+$22] Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. *arXiv preprint arXiv:2202.05791*, 2022. 3

[HK14] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014. 2

[JZ13] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013. 2

[RRWN11] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *Advances in neural information processing systems*, 24, 2011. 2

[SLRB17] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, 2017. 2

[SR13] Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013. 2

[SSZ14] Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *International conference on machine learning*, pages 64–72. PMLR, 2014. 1

[ZDS$^+$19] Kaiwen Zhou, Qinghua Ding, Fanhua Shang, James Cheng, Danli Li, and Zhi-Quan Luo. Direct acceleration of saga using sampled negative momentum. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1602–1610. PMLR, 2019. 1

A Preliminaries

Lemma 2. *Suppose f is μ-strongly convex and L-smooth. Let $x^* = \operatorname{argmin}_x f(x)$, then:*

$$f(x) - f(x^*) \leq \frac{1}{2\mu} \left\|\nabla f(x)\right\|^2 \leq \frac{L^2}{\mu^2}(f(x) - f(x^*)) \tag{17}$$

Proof. By definition of μ-strongly convex function,

$$\begin{aligned} f(x^*) &\geq f(x) + \langle \nabla f(x), x^* - x\rangle + \frac{\mu}{2}\left\|x^* - x\right\|^2 \\ \implies f(x) - f(x^*) &\leq \langle \nabla f(x), x - x^*\rangle - \frac{\mu}{2}\left\|x^* - x\right\|^2 \\ &= -\frac{\mu}{2}\left(\left\|x^* - x\right\|^2 + \frac{2}{\mu}\langle \nabla f(x), x^* - x\rangle + \frac{1}{\mu^2}\left\|\nabla f(x)\right\|^2 - \frac{1}{\mu^2}\left\|\nabla f(x)\right\|^2\right) \\ &= \frac{1}{2\mu}\left\|\nabla f(x)\right\|^2 - \frac{\mu}{2}\left\|x - \frac{1}{\mu}\nabla f(x) - x^*\right\|^2 \leq \frac{1}{2\mu}\left\|\nabla f(x)\right\|^2 \end{aligned}$$

To get the upper bound, observe

$$\left\|\nabla f(x)\right\|^2 = \left\|\nabla f(x) - \nabla f(x^*)\right\|^2 \leq L^2\left\|x - x^*\right\|^2 \tag{18}$$

because $\nabla f(x^*) = 0$. From strong convexity,

$$\begin{aligned} f(x) &\geq f(x^*) + \langle \nabla f(x^*), x - x^*\rangle + \frac{\mu}{2}\left\|x^* - x\right\|^2 \\ \implies \left\|x - x^*\right\|^2 &\leq \frac{2}{\mu}\left(f(x) - f(x^*)\right) \end{aligned} \tag{19}$$

Combining (18) and (19) completes the proof.

□

B Proof of Lemma 1

Proof. By definition of L-smooth function,

$$\begin{aligned} f(X_{k,j+1}) &\leq f(X_{k,j}) + \langle \nabla f(X_{k,j}), X_{k,j+1} - X_{k,j}\rangle + \frac{L}{2}\left\|X_{k,j+1} - X_{k,j}\right\|^2 \\ &= f(X_{k,j}) - \eta_k\langle \nabla f(X_{k,j}), \nabla F(X_{k,j}, \xi'_j) - (\nabla F(X_k, \xi'_j) - G(X_k))\rangle \\ &\quad + \frac{L}{2}\eta_k^2\left\|\nabla F(X_{k,j}, \xi'_j) - (\nabla F(X_k, \xi'_j) - G(X_k))\right\|^2 \end{aligned}$$

Taking expectation conditioned on $\mathcal{F}_k^j$:

$$\begin{aligned} \mathbb{E}\left[f(X_{k,j+1})|\mathcal{F}_k^j\right] &= f(X_{k,j}) - \eta_k\left\langle \nabla f(X_{k,j}), \mathbb{E}\left[\nabla F(X_{k,j}, \xi'_j) - (\nabla F(X_k, \xi'_j) - G(X_k))|\mathcal{F}_k^j\right]\right\rangle \\ &\quad + \frac{L}{2}\eta_k^2\mathbb{E}\left[\left\|\nabla F(X_{k,j}, \xi'_j) - (\nabla F(X_k, \xi'_j) - G(X_k))\right\|^2|\mathcal{F}_k^j\right] \\ &= f(X_{k,j}) - \eta_k\left\|\nabla f(X_{k,j})\right\|^2 + \eta_k\left\langle \nabla f(X_{k,j}), \nabla f(X_k) - G(X_k)\right\rangle \\ &\quad + \frac{L}{2}\eta_k^2\mathbb{E}\left[\left\|\nabla F(X_{k,j}, \xi'_j) - (\nabla F(X_k, \xi'_j) - G(X_k))\right\|^2|\mathcal{F}_k^j\right] \end{aligned} \tag{20}$$

Consider:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla F(X_{k,j},\xi_j') - (\nabla F(X_k,\xi_j') - G(X_k))\right\|^2 |\mathcal{F}_k^j\right] \\
&\overset{(a)}{\le} 4\mathbb{E}\left[\left\|\nabla F(X_{k,j},\xi_j')\right\|^2 |\mathcal{F}_k^j\right] + 4\mathbb{E}\left[\left\|\nabla F(X_k,\xi_j')\right\|^2 |\mathcal{F}_k^j\right] + 2\mathbb{E}\left[\left\|G(X_k)\right\|^2 |\mathcal{F}_k^j\right] \\
&\overset{(b)}{=} 4\mathbb{E}\left[\left\|\nabla F(X_{k,j},\xi_j')\right\|^2 |\mathcal{F}_k^j\right] + 4\mathbb{E}\left[\left\|\nabla F(X_k,\xi_j')\right\|^2 |\mathcal{F}_k^j\right] + 2\left\|G(X_k)\right\|^2 \\
&\overset{(3),(4)}{\le} 4\left(2\sigma_0^2 + (1+\sigma_1^2)(\|\nabla f(X_{k,j})\|^2 + \|\nabla f(X_k)\|^2)\right) + 2\left\|G(X_k)\right\|^2
\end{aligned} \tag{21}$$

where (a) holds by applying $\|a+b\|^2 \le 2(\|a\|^2 + \|b\|^2)$ twice,
(b) holds because the $\mathcal{F}_k^j$ is independent of the randomness in $G(X_k)$.

Substituting (21) in (20),

$$\begin{aligned}
\mathbb{E}\left[f(X_{k,j+1})|\mathcal{F}_k^j\right] &= f(X_{k,j}) - \eta_k \left\|\nabla f(X_{k,j})\right\|^2 + \eta_k \left\langle \nabla f(X_{k,j}), \nabla f(X_k) - G(X_k)\right\rangle \\
&\quad + L\eta_k^2\left(4\sigma_0^2 + 2(1+\sigma_1^2)(\|\nabla f(X_{k,j})\|^2 + \|\nabla f(X_k)\|^2) + \|G(X_k)\|^2\right) \\
&= f(X_{k,j}) - \eta_k(1 - 2\eta_k L(1+\sigma_1^2)) \left\|\nabla f(X_{k,j})\right\|^2 + \eta_k \left\langle \nabla f(X_{k,j}), \nabla f(X_k) - G(X_k)\right\rangle \\
&\quad + L\eta_k^2\left(4\sigma_0^2 + 2(1+\sigma_1^2) \left\|\nabla f(X_k)\right\|^2 + \left\|G(X_k)\right\|^2\right) \\
&\overset{(a)}{\le} f(X_{k,j}) - 2\mu\eta_k(1 - 2\eta_k L(1+\sigma_1^2))(f(X_{k,j}) - f(x^*)) + \eta_k \left\langle \nabla f(X_{k,j}), \nabla f(X_k) - G(X_k)\right\rangle \\
&\quad + L\eta_k^2\left(4\sigma_0^2 + 2(1+\sigma_1^2) \left\|\nabla f(X_k)\right\|^2 + \left\|G(X_k)\right\|^2\right)
\end{aligned}$$

where (a) follows from Lemma 2.

Taking expectations w.r.t $\xi_1, \xi_2, \ldots, \xi_{n_k}$:

$$\begin{aligned}
\mathbb{E}\left[f(X_{k,j+1})|\mathcal{F}_k^j\right] &\le f(X_{k,j}) - 2\mu\eta_k(1 - 2\eta_k L(1+\sigma_1^2))(f(X_{k,j}) - f(x^*)) \\
&\quad + \eta_k \left\langle \nabla f(X_{k,j}), \mathbb{E}\left[\nabla f(X_k) - G(X_k)\right]\right\rangle \\
&\quad + L\eta_k^2\left(4\sigma_0^2 + 2(1+\sigma_1^2) \left\|\nabla f(X_k)\right\|^2 + \mathbb{E}\left[\left\|G(X_k)\right\|^2\right]\right) \\
&\overset{(a)}{=} f(X_{k,j}) - 2\mu\eta_k(1 - 2\eta_k L(1+\sigma_1^2))(f(X_{k,j}) - f(x^*)) \\
&\quad + L\eta_k^2\left(\left(4 + \frac{1}{n_k}\right)\sigma_0^2 + \left(3 + 2\sigma_1^2 + \frac{\sigma_1^2}{n_k}\right)\left\|\nabla f(X_k)\right\|^2\right) \qquad (22) \\
&\overset{(b)}{\le} f(X_{k,j}) - 2\mu\eta_k(1 - 2\eta_k L(1+\sigma_1^2))(f(X_{k,j}) - f(x^*)) \\
&\quad + L\eta_k^2\left(\left(4 + \frac{1}{n_k}\right)\sigma_0^2 + \frac{2L^2}{\mu}\left(3 + 2\sigma_1^2 + \frac{\sigma_1^2}{n_k}\right)(f(X_k) - f(x^*))\right) \qquad (23)
\end{aligned}$$

where (a) is true because:

$$\begin{aligned}
\mathbb{E}\left[\left\|G(X_k)\right\|^2\right] &= \mathbb{E}\left[\left\|G(X_k) - \nabla f(X_k) + \nabla f(X_k)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|G(X_k) - \nabla f(X_k)\right\|^2\right] + \left\|\nabla f(X_k)\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left\|\frac{1}{n_k}\sum_{j=1}^{n_k}\left(\nabla F(X_k,\xi_j) - \nabla f(X_k)\right)\right\|^2\right] + \|\nabla f(X_k)\|^2 \\
&\leq \frac{1}{n_k}\left(\sigma_0^2 + \sigma_1^2\|\nabla f(X_k)\|\right) + \|\nabla f(X_k)\|^2
\end{aligned}
$$

and (b) follows from Lemma 2.

Finally, taking expectations, subtracting $f(x^*)$ on both sides and denoting $\Delta_t = \mathbb{E}\left[f(X_t)\right] - f(x^*)$,

$$
\Delta_{k,j+1} \leq \left(1 - 2\mu\eta_k(1 - 2\eta_k L(1+\sigma_1^2)\right)\Delta_{k,j} + L\eta_k^2\left(\left(4 + \frac{1}{n_k}\right)\sigma_0^2 + \frac{2L^2}{\mu}\left(3 + 2\sigma_1^2 + \frac{\sigma_1^2}{n_k}\right)\Delta_k\right) \tag{24}
$$

□

C Proof of Theorem 1

Proof. Observe that

$$
\eta_k = \eta \leq \frac{\mu^2}{12L^3(1+\sigma_1^2)} < \frac{1}{4L(1+\sigma_1^2)}
$$

So, using Lemma 1, we have:

$$
\Delta_{k,j+1} \leq \left(1 - 2\mu\eta_k(1 - 2\eta_k L(1+\sigma_1^2)\right)\Delta_{k,j} + L\eta_k^2\left(\left(4 + \frac{1}{n_k}\right)\sigma_0^2 + \frac{2L^2}{\mu}\left(3 + 2\sigma_1^2 + \frac{\sigma_1^2}{n_k}\right)\Delta_k\right) \tag{25}
$$

Furthermore, define

$$
\begin{aligned}
\alpha &:= \left(1 - 2\mu\eta(1 - 2\eta L(1+\sigma_1^2)\right) \\
\beta_k &:= \left(4 + \frac{1}{n_k}\right)L\eta_k^2\sigma_0^2 \\
\gamma_k &:= \frac{2L^3\eta_k^2}{\mu}\left(3 + 2\sigma_1^2 + \frac{\sigma_1^2}{n_k}\right)\Delta_k
\end{aligned}
$$

It is easy to verify $0 < \alpha < 1$. So,

$$
\begin{aligned}
& \Delta_{k,j+1} \leq \alpha\Delta_{k,j} + \beta_k + \gamma_k \\
\Longrightarrow\ & \Delta_{k,j+1} \leq \alpha(\alpha\Delta_{k,j-1} + \beta_k + \gamma_k) + \beta_k + \gamma_k \\
&\qquad = \alpha^2\Delta_{k,j-1} + (\beta_k + \gamma_k)(1+\alpha)
\end{aligned}
$$

Setting $j = m$ and continuing the recursion till $j = 1$, , and substituting $\Delta_{k,m+1} = \Delta_{k+1}$ and $\Delta_{k,1} = \Delta_k$

$$
\begin{aligned}
\Delta_{k+1} &\leq \alpha^m\Delta_k + \frac{\beta_k + \gamma_k}{1-\alpha} \\
&= \left[\alpha^m + \frac{2L^3\eta^2}{\mu(1-\alpha)}\left(3 + 2\sigma_1^2 + \frac{\sigma_1^2}{n_k}\right)\right]\Delta_k + \frac{\beta_k}{1-\alpha} \\
&\leq \rho\Delta_k + \frac{\beta_k}{1-\alpha}
\end{aligned}
$$

where

$$\rho = \frac{1}{1-2\eta L(1+\sigma_1^2)}\left(\frac{1}{1+2m\mu\eta} + \frac{3L^3\eta(1+\sigma_1^2)}{\mu^2}\right) \tag{26}$$

We now show that $0 < \rho < 1$. Note that the lower bound is trivial.

$$\begin{aligned}
\alpha^m + \frac{2L}{\mu(1-\alpha)}\left(3 + 2\sigma_1^2 + \frac{\sigma_1^2}{n_k}\right) &\overset{(a)}{\leq} \alpha^m + \frac{6L^3\eta^2(1+\sigma_1^2)}{\mu(1-\alpha)} \\
&= \left(1 - 2\mu\eta(1-2\eta L(1+\sigma_1^2)\right)^m + \frac{3L^3\eta(1+\sigma_1^2)}{\mu^2(1-2\eta L(1+\sigma_1^2))} \\
&\overset{(b)}{\leq} \frac{1}{1+2m\mu\eta(1-2\eta L(1+\sigma_1^2))} + \frac{3L^3\eta(1+\sigma_1^2)}{\mu^2(1-2\eta L(1+\sigma_1^2))} \\
&\overset{(c)}{\leq} \frac{1}{1-2\eta L(1+\sigma_1^2)}\left(\frac{1}{1+2m\mu\eta} + \frac{3L^3\eta(1+\sigma_1^2)}{\mu^2}\right) = \rho
\end{aligned}$$

where (a) follows because $n_k \geq 1$
(b) follows as for $x \in (0,1), n \geq 1$, $(1-x)^n \leq \dfrac{1}{1+nx}$.
(c) follows because for $a > 0, b \in (0,1)$, $\dfrac{1}{1+ab} \leq \dfrac{1}{b(1+a)}$.

Now, observe:

$$\eta \leq \frac{\mu^2}{12L^3(1+\sigma_1^2)} < \frac{1}{4L(1+\sigma_1^2)} \implies \frac{1}{1-2\eta L(1+\sigma_1^2)} < 2 \tag{27}$$

$$m \geq \frac{3}{2\mu\eta} \implies \frac{1}{1+2m\mu\eta} \leq \frac{1}{4} \tag{28}$$

$$\eta \leq \frac{\mu^2}{12L^3(1+\sigma_1^2)} \implies \frac{3L^3\eta(1+\sigma_1^2)}{\mu^2} \leq \frac{1}{4} \tag{29}$$

Combining the above results, we get $\rho < 1$.

Performing a similar recursion for the outer loop,

$$\begin{aligned}
\Delta_{k+1} \leq \rho\Delta_k + \frac{\beta_k}{1-\alpha} &\leq \rho\left(\rho\Delta_{k-1} + \frac{\beta_{k-1}}{1-\alpha}\right) + \frac{\beta_k}{1-\alpha} \\
&= \rho^2\Delta_{k-1} + \frac{1}{1-\alpha}\left(\rho\beta_{k-1} + \beta_k\right) \\
\implies \Delta_{k+1} &\leq \rho^k\Delta_1 + \frac{1}{1-\alpha}\sum_{r=1}^{k}\rho^r\beta_{k-r}
\end{aligned} \tag{30}$$

C.1 Growth Condition

By definition of GC, $\sigma_0 = 0 \implies \beta_t = 0, \forall t \geq 1$. So, we obtain *linear* convergence, i.e.,

$$\Delta_{k+1} \leq \rho^k\Delta_1 \tag{31}$$

Complexity Analysis:

Observe that $\log(\frac{1}{\rho}) \leq 1 - \rho$, for $\rho \in (0, 1)$. So,

$$
\begin{aligned}
k \geq \frac{1}{1-\rho} \log\left(\frac{\Delta_1}{\varepsilon}\right) &\implies k \log\left(\frac{1}{\rho}\right) \geq \log\left(\frac{\Delta_1}{\varepsilon}\right) \\
&\implies \Delta_{k+1} \overset{(31)}{\leq} \rho^k \Delta_1 \leq \varepsilon.
\end{aligned} \tag{32}
$$

C.2 Weak Growth Condition

Observe that *irrespective of the way* n_k varies,

$$
\beta_k = \left(4 + \frac{1}{n_k}\right) L\eta^2\sigma_0^2 \leq 5L\eta^2\sigma_0^2 \tag{33}
$$

as $n_k \geq 1$. So,

$$
\begin{aligned}
\Delta_{k+1} &\leq \rho^k \Delta_1 + \frac{1}{1-\alpha} \sum_{r=1}^{k} \rho^r \beta_{k-r} \\
&\leq \rho^k \Delta_1 + \frac{5L\eta^2\sigma_0^2}{(1-\alpha)(1-\rho)} \\
&\leq \rho^k \Delta_1 + \frac{5L\eta\sigma_0^2}{2\mu(1-\rho)(1-2\eta L(1+\sigma_1^2))} \\
&\overset{(27)}{\leq} \rho^k \Delta_1 + \frac{5L\eta\sigma_0^2}{\mu(1-\rho)}
\end{aligned}
$$

Hence, in this case, the algorithm converges to the proximity of the optimal value *linearly*, however, it does not converge exactly as $k \to \infty$. Instead, we get a residual term. Precisely,

$$
\Delta_\infty \lesssim \frac{L\eta\sigma_0^2}{\mu(1-\rho)} \tag{34}
$$

Complexity Analysis:

We intend to upper bound each of the individual terms by $\varepsilon/2$. Observe that $\log(\frac{1}{\rho}) \leq 1 - \rho$, for $\rho \in (0, 1)$. So,

$$
k \geq \frac{1}{1-\rho} \log\left(\frac{2\Delta_1}{\varepsilon}\right) \implies k \log\left(\frac{1}{\rho}\right) \geq \log\left(\frac{2\Delta_1}{\varepsilon}\right) \implies \rho^k \Delta_1 \leq \frac{\varepsilon}{2}. \tag{35}
$$

Bounding the second term:

$$
\eta \overset{(10)}{\leq} \frac{\varepsilon\mu}{30L\sigma_0^2} \implies \frac{5L\eta\sigma_0^2}{\mu(1-\rho)} \leq \frac{\varepsilon}{6(1-\rho)} \tag{36}
$$

By definition of ρ, (28) and (29), observe that

$$
\begin{aligned}
\rho &\leq \frac{1}{2(1-2\eta L(1+\sigma_1^2))} \\
\implies \frac{1}{1-\rho} &\leq \frac{2(1-2\eta L(1+\sigma_1^2))}{(1-4\eta L(1+\sigma_1^2))} \leq \frac{2}{(1-4\eta L(1+\sigma_1^2))}
\end{aligned} \tag{37}
$$

Now,

$$\eta \overset{(10)}{\leq} \frac{\mu^2}{12L^3(1+\sigma_1^2)} \implies \frac{1}{(1-4\eta L(1+\sigma_1^2))} \leq \left(1-\frac{\mu^2}{3L^2}\right)^{-1} \leq \frac{3}{2} \tag{38}$$

where the last inequality holds because $L \geq \mu \geq 0$. So,

$$\frac{5L\eta\sigma_0^2}{\mu(1-\rho)} \overset{(36)}{\leq} \frac{\varepsilon}{6(1-\rho)} \overset{(37)}{\leq} \frac{\varepsilon}{3(1-4\eta L(1+\sigma_1^2))} \overset{(38)}{\leq} \frac{\varepsilon}{2}. \tag{39}$$

Combining (35) and (39), we get the desired result.

□

D Proof of Theorem 2

Proof. Observe

$$\eta_k < \eta_1 \leq \frac{\mu^2}{12L^3(1+\sigma_1^2)} < \frac{1}{4L(1+\sigma_1^2)} \tag{40}$$

So, using Lemma 1:

$$\begin{aligned}\Delta_{k,j+1} &\leq \left(1-2\mu\eta_k(1-2\eta_k L(1+\sigma_1^2)\right)\Delta_{k,j} + L\eta_k^2\left(\left(4+\frac{1}{n_k}\right)\sigma_0^2 + \frac{2L^2}{\mu}\left(3+2\sigma_1^2+\frac{\sigma_1^2}{n_k}\right)\Delta_k\right) \\ &\leq \left(1-2\mu\eta_k(1-2\eta_k L(1+\sigma_1^2)\right)\Delta_{k,j} + L\eta_k^2\left(5\sigma_0^2 + \frac{6L^2(1+\sigma_1^2)}{\mu}\Delta_k\right)\end{aligned}$$

Furthermore, define

$$\begin{aligned}\alpha_k &:= 1-2\mu\eta_k(1-2\eta_k L(1+\sigma_1^2)) \\ \beta_k &:= 5L\eta_k^2\sigma_0^2 \\ \gamma_k &:= \frac{6L^3\eta_k^2(1+\sigma_1^2)}{\mu}\Delta_k\end{aligned}$$

Doing a recursion from $j = m$ to $j = 1$, and substituting $\Delta_{k,m+1} = \Delta_{k+1}$ and $\Delta_{k,1} = \Delta_k$:

$$\Delta_{k+1} \leq \alpha_k^m \Delta_k + \frac{(\beta_k+\gamma_k)(1-\alpha_k^m)}{1-\alpha_k} \tag{41}$$

Observe that $\eta_k \leq \eta_1 \leq \frac{\mu^2}{12L^3(1+\sigma_1^2)} \implies \alpha_k \leq 1-2\mu\eta_k(1-\frac{\mu^2}{6L^2}) \leq 1-\frac{5}{3}\mu\eta_k$. Furthermore,

$$\begin{aligned}\alpha_k &= 1-2\mu\eta_k(1-2\eta_k L(1+\sigma_1^2)) \\ &= 1-2\mu\eta_k+\mu^2\eta_k^2-\mu^2\eta_k^2+4\mu\eta_k^2 L(1+\sigma_1^2) \\ &= (1-\mu\eta_k)^2+\mu\eta_k^2(4L(1+\sigma_1^2)-\mu) \geq (1-\mu\eta_k)^2\end{aligned}$$

The last inequality follows because by definition, $L \geq \mu \geq 0$. So,

$$(1-\mu\eta_k)^2 \leq \alpha_k \leq 1-\frac{5}{3}\mu\eta_k \tag{42}$$

Hence,

$$\begin{aligned}
\Delta_{k+1} &\leq \alpha_k^m \Delta_k + \frac{\beta_k(1-\alpha_k^m)}{1-\alpha_k} \\
&\overset{(42)}{\leq} \left(1 - \frac{5}{3}\mu\eta_k\right)^m \Delta_k + \frac{3(\beta_k + \gamma_k)(1-(1-\mu\eta_k)^{2m})}{5\mu\eta_k} \\
&\overset{(a)}{\leq} \left(1 - \frac{5}{3}\mu\eta_k\right)^m \Delta_k + \frac{6}{5}m(\beta_k + \gamma_k) \\
&= \left(1 - \frac{5}{3}\mu\eta_k\right)^m \Delta_k + 6mL\eta_k^2\sigma_0^2 + \frac{36mL^3\eta_k^2(1+\sigma_1^2)}{5\mu}\Delta_k \\
&\leq \left(1 - \frac{5}{3}\mu\eta_k\right)^m \Delta_k + 6mL\eta_k^2\sigma_0^2 + \frac{3m\mu\eta_k}{5}\Delta_k \\
&\overset{(b)}{\leq} \left(\frac{1}{1+\frac{5m\mu\eta_k}{3}} + \frac{3m\mu\eta_k}{5}\right)\Delta_k + 6mL\eta_k^2\sigma_0^2 \\
&\overset{(c)}{\leq} \left(1 - \frac{2}{5}m\mu\eta_k\right)\Delta_k + 6mL\eta_k^2\sigma_0^2
\end{aligned}$$

where (a) and (b) follows as $1 - mx \leq (1-x)^m \leq \min\{1-x, \frac{1}{1+mx}\}$ for $x \in (0,1)$ and $m \geq 1$. (c) follows because by choice of m, $m\mu\eta_k \leq \frac{2}{5}$, and for $x \in (0, \frac{2}{5})$,

$$\frac{1}{1+\frac{5x}{3}} + \frac{3x}{5} \leq 1 - \frac{2x}{5}$$

Hence, we have:

$$\Delta_{k+1} \leq \left(1 - \frac{2}{5}m\mu\eta_k\right)\Delta_k + 6mL\eta_k^2\sigma_0^2 \tag{43}$$

D.1 Growth Condition

If $\sigma_0 = 0$, then,

$$\begin{aligned}
\Delta_{k+1} &\leq \left(1 - \frac{2}{5}m\mu\eta_k\right)\Delta_k \\
&\leq \exp\left(-\frac{2}{5}m\mu\eta_k\right)\Delta_k \\
\implies \Delta_{k+1} &\leq \exp\left(-\frac{2}{5}m\mu\sum_{j=1}^{k}\eta_j\right)\Delta_1
\end{aligned}$$

where the second inequality follows as $1 - x \leq e^{-x}$.

$$\begin{aligned}
\Delta_{k+1} &\leq \exp\left(-\frac{2}{5}m\mu\sum_{j=1}^{k}\eta_j\right)\Delta_1 \\
&\overset{(12)}{=} \exp\left(-\frac{2}{5}m\mu\theta\sum_{j=1}^{k}\frac{1}{\lambda+j}\right)\Delta_1 \\
&\overset{(a)}{\leq} \exp\left(-\frac{2}{5}m\mu\theta\log\left(\frac{\lambda+k}{\lambda+1}\right)\right)\Delta_1 \\
&= \left(\frac{\lambda+1}{\lambda+k}\right)^{c}\Delta_1
\end{aligned}$$

where $c = \dfrac{2}{5}m\mu\theta \overset{(12)}{>} 1$.

Complexity Analysis:

Rearranging the terms in the lower bound of k:

$$k \geq \left(\frac{\Delta_1}{\varepsilon}\right)^{1/c}(\lambda+1) - \lambda \implies \left(\frac{\lambda+1}{\lambda+k}\right)^{c}\Delta_1 \leq \varepsilon \implies \Delta_{k+1} \leq \varepsilon$$

which completes the proof.

D.2 Weak Growth Condition

Similar to [BCN18], we prove the result following the principle of mathematical induction on k. For $k = 1$, the result is obvious from the definition of ν. Suppose the result holds true for some $k \geq 1$. Then, from (43),

$$\begin{aligned}
\Delta_{k+1} &\leq \left(1 - \frac{2}{5}m\mu\eta_k\right)\Delta_k + 6mL\eta_k^2\sigma_0^2 \\
&\leq \left(1 - \frac{2}{5}m\mu\eta_k\right)\frac{\nu}{\lambda+k} + 6mL\eta_k^2\sigma_0^2 \\
&\overset{(12)}{\leq} \left(1 - \frac{2m\mu\theta/5}{\lambda+k}\right)\frac{\nu}{\lambda+k} + \frac{6mL\theta^2\sigma_0^2}{(\lambda+k)^2} \\
&= \frac{(\lambda+k-1)\nu}{(\lambda+k)^2} - \frac{(2m\mu\theta/5-1)\nu}{(\lambda+k)^2} + \frac{6mL\theta^2\sigma_0^2}{(\lambda+k)^2} \\
&\overset{(a)}{\leq} \frac{(\lambda+k-1)\nu}{(\lambda+k)^2} \overset{(b)}{\leq} \frac{\nu}{\lambda+k+1}
\end{aligned}$$

where (a) follows from the definition of ν and (b) holds true as $(z+1)(z-1) \leq z^2$.

Complexity Analysis:

Rearranging the terms in the lower bound of k:

$$k \geq \frac{\nu}{\varepsilon} - \lambda \implies \frac{\nu}{k+\lambda} \leq \varepsilon \implies \Delta_{k+1} \leq \varepsilon$$

which completes the proof.

□