Fast Convergence Rates of Random Reshuffling under Interpolation

author names withheld

Under Review for OPT 2022

Abstract

Modern machine learning models are highly over-parameterized and as a result they can interpolate the training data. Under such scenario, we study the convergence properties of a sampling-without-replacement variant of Stochastic Gradient Descent (SGD), known as Random Reshuffling. Unlike SGD that samples data with replacement at every iteration, Random Reshuffling chooses a random permutation of data at the beginning of each epoch. It has recently been shown that Random Reshuffling converges faster than SGD when the number of epochs is larger than the condition number (κ) of the problem. Under over-parameterization, we show that instead of requiring such large number of epochs, Random Reshuffling performs better than SGD when the number of samples is less than κ, and this holds true for strongly-convex, convex, and non-convex objectives. For strongly-convex objectives, our analysis does not require each individual function to be convex or Lipschitz, relaxing common assumptions made in previous works. Moreover, we show that Random Reshuffling outperforms Incremental Gradient for either large or small κ under interpolation. Our empirical results support our theoretical conclusions.

1. Introduction

We consider finite-sum minimization problems of the form

$$\min\left\{ f(x) = \frac{1}{n}\sum_{i=1}^{n} f(x;i) \right\}. \tag{1}$$

Stochastic gradient descent (SGD) is a popular algorithm for solving machine learning problems of this form. As such, in recent years, a significant amount of effort has been made to understand its theoretical and empirical properties (Bottou et al., 2018). SGD has a simple update rule in which a sample i_k is chosen randomly with replacement at each iteration to compute $x^{k+1} = x^k - \eta^k \nabla f(x^k; i_k)$. This is cheaper than using the full gradient at each iteration. However, it is well known that the convergence rates of SGD, i.e. $\mathcal{O}(\frac{1}{k})$ and $\mathcal{O}(\frac{1}{\sqrt{k}})$ for strongly-convex and convex objectives respectively (Nemirovski et al., 2009), are worse than those of full gradient descent.

Given the increasing complexity of modern learning models, a practically-relevant question to ask is how SGD performs in over-parameterized settings, under which the model fits or interpolates the data completely. Previously, it has been shown that SGD can achieve the rate of full gradient descent under a strong growth condition (SGC) which implies interpolation (Schmidt and Roux, 2013). Besides this, a sampling-without-replacement counterpart of SGD, known as Random Reshuffling, has been shown to converge faster than SGD empirically (Bottou, 2012). However, its analysis is substantially more difficult than SGD and has only emerged in a series of recent efforts because each individual gradient is no longer an unbiased estimate of the full gradient conditioned on the

© .

past iterates (Gürbüzbalaban et al., 2021, Haochen and Sra, 2019, Nagaraj et al., 2019, Safran and Shamir, 2020). One important theoretical aspect of Random Reshuffling that still remains elusive is its convergence properties under interpolation, which is one of the most common situations considering its frequent use in training over-parameterized models such as a deep neural network. In this work, we aim to address this question with our theory and experiments.

Table 1: Number of gradient evaluations required by each algorithm to obtain an ϵ-accurate solution, which is defined as $\|x - x^*\|^2 \leq \epsilon$, $f(x) - f(x^*) \leq \epsilon$, and $\|\nabla f(x)\| \leq \epsilon$ for μ-strongly convex, convex, and non-convex objectives respectively.[1] Our results hold under SGC (Assumption 2). See text or Table 2 in Appendix A for bounds under a relaxed SGC condition (Assumption 3).

Citation	Algorithm	μ-Strongly Convex	Convex	Non-Convex
Mishchenko et al. (2020)[2]	IG	$\kappa n + \frac{n\sqrt{\kappa}\sigma^*}{\mu\sqrt{\epsilon}}$	$\frac{Ln}{\epsilon} + \frac{\sqrt{L}n\sigma^*}{\epsilon^{3/2}}$	$\frac{Ln}{\epsilon^2} + \frac{Ln(\sqrt{A}+B)}{\epsilon^3}$
	RR	$\kappa n + \frac{\sqrt{n}\kappa\sigma^*}{\mu\sqrt{\epsilon}}$	$\frac{Ln}{\epsilon} + \frac{\sqrt{L}n\sigma^*}{\epsilon^{3/2}}$	$\frac{Ln}{\epsilon^2} + \frac{L\sqrt{n}(\sqrt{A}+B)}{\epsilon^3}$
Nguyen et al. (2021)[3]	IG	$\frac{n}{\sqrt{\mu\epsilon}} + \frac{(n\kappa+1)\sigma^*}{\mu\sqrt{\epsilon}}$	-	$\frac{Ln\sqrt{3\Theta+2}}{\epsilon^2} + \frac{Ln\sigma}{\epsilon^3}$
	RR	$\frac{\sqrt{n}}{\sqrt{\mu\epsilon}} + \frac{\sqrt{n}\kappa\sigma}{\mu\sqrt{\epsilon}}$	-	$\frac{L\sqrt{n}\sqrt{\Theta+n}}{\epsilon^2} + \frac{L\sqrt{n}\sigma}{\epsilon^3}$
Vaswani et al. (2019)	SGD	$\kappa\rho$	$\frac{L\rho+\rho-1}{\epsilon}$	$\frac{L\rho}{\epsilon^2}$
This work	IG	$\kappa n\sqrt{\rho}$	$\frac{Ln\sqrt{\rho}}{\epsilon}$	$\frac{Ln\sqrt{\rho}}{\epsilon^2}$
	RR	$\kappa\sqrt{n}\sqrt{\rho+1+n}$	$\frac{L\sqrt{n}\sqrt{\rho+1+n}}{\epsilon}$	$\frac{L\sqrt{n}\sqrt{\rho+1+n}}{\epsilon^2}$

[1] We ignore numerical constants and logarithmic factors.

[2] Mishchenko et al. (2020) assume each f_i to be convex for μ-strongly convex objectives. We do not make this assumption. For non-convex objectives, they assume $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x;i) - \nabla f(x)\|^2 \leq 2A(f(x) - f^*) + B^2$, and denote $\sigma^* = \frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x^*;i)\|^2$.

[3] Nguyen et al. (2021) make the assumption $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x;i) - \nabla f(x)\|^2 \leq \Theta\|\nabla f(x)\|^2 + \sigma^2$ for analyzing RR. Note that this assumption is equivalent to Assumption 3 with $\Theta = \rho - 1$. Despite that, our refined analysis results in a tighter bound for μ-strongly convex objectives (see Table 2 in Appendix A). We remark that their results for IG hold for more general shuffling schemes.

1.1. Problem Statement

The update rule of Random Reshuffling (RR) is as follows:

$$x_{j+1}^t = x_j^t - \eta_j^t \nabla f(x_j^t; \pi_{j+1}^t). \tag{2}$$

That is, at each epoch $t \in \{1, 2, ..., T\}$, π^t is chosen randomly from the set of all possible permutations, and π_{j+1}^t is sampled without replacement from the set $\{1, 2, ..., n\}$ for $j \in \{0, 1, ..., n-1\}$. Note that $x_0 \triangleq x_0^1$ is the initialization and $x_0^{t+1} = x_n^t \ \forall t \geq 1$. A deterministic counterpart of RR is Incremental Gradient method (IG), where π^t is deterministic and fixed over all epochs. For IG, we take the ordering to be cyclic, i.e. $\pi_{j+1}^t = j + 1$. We study RR and IG under interpolation (implied by the SGC) for strongly-convex, convex, and non-convex objectives.

1.2. Assumptions

In this section, we present the assumptions made in our analysis. Assumption 1 is used to obtain all the results.

Assumption 1 *Each individual loss $f(\cdot; i)$ is L-smooth, which means the objective f is also L-smooth*

$$\|\nabla f(x; i) - \nabla f(x'; i)\| \leq L\|x - x'\|, \quad \forall x, x' \in dom(f). \tag{3}$$

We also assume f is lower bounded by f^*, which is achieved at some x^*, so $f^* = f(x^*)$.

Assumption 2 (SGC) *There exists a constant $\rho \geq 1$ such that the following holds*

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x; i)\|^2 \leq \rho\|\nabla f(x)\|^2, \quad \forall x \in dom(f). \tag{4}$$

Note that the SGC necessarily implies interpolation, which is defined as $\nabla f(x^*; i) = 0$ for $\forall i$ where x^* satisfies $\nabla f(x^*) = 0$ (Vaswani et al., 2019). We refer the reader to Vaswani et al. (2019) for examples of settings in which SGC holds. For example, a smooth objective f that satisfies the Polyak-Łojasiewicz (PL) inequality satisfies SGC with $\rho = \frac{L}{\mu}$ under interpolation. Another example is the squared hinge loss under separable data.

Assumption 3 *There exists constants $\rho \geq 0$ and $\sigma \geq 0$ such that the following holds*

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x; i)\|^2 \leq \rho\|\nabla f(x)\|^2 + \sigma^2, \quad \forall x \in dom(f). \tag{5}$$

Note that Assumption 3 reduces to the SGC when $\sigma = 0$, and reduces to the bounded variance assumption $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x; i) - \nabla f(x)\|^2 \leq \sigma^2$ when $\rho = 1$, which is commonly used in the analysis of SGD (Bottou et al., 2018). We use Assumptions 1 and 3 when analyzing strongly-convex and non-convex objectives.

1.3. Related Work

The non-asymptotic convergence properties of RR have only been addressed recently. Haochen and Sra (2019) give the first convergence result of $\tilde{\mathcal{O}}(\frac{1}{n^2T^2} + \frac{1}{T^3})$. They also show that RR is at least as fast as SGD under interpolation (Haochen and Sra, 2019). Recall that T is the number of epochs. This rate is improved by Nagaraj et al. (2019) to $\tilde{\mathcal{O}}(\frac{1}{nT^2})$ when $T \gtrsim \kappa^2$ by exploiting the component-wise convexity of each $f(\cdot; i)$, with a matching lower bound of $\Omega(\frac{1}{nT^2})$ given by Rajput et al. (2020). Note that this rate is faster than the $\tilde{\mathcal{O}}(\frac{1}{nT})$ rate of SGD when the large epoch requirement is satisfied. Mishchenko et al. (2020) obtain the same rate of $\tilde{\mathcal{O}}(\frac{1}{nT^2})$ but only require $T \gtrsim \kappa$. Their analysis is also dependent on the underlying component-wise convexity structure. Ahn et al. (2020) remove this dependence and obtain the same rate with $T \gtrsim \kappa$. However, their analysis relies on each $f(\cdot; i)$ being G-Lipschitz ($\|\nabla f(\cdot; i)\| \leq G$ for $\forall i$), which implicitly places a ball constraint on problem (1) and a projection operation is needed to ensure iterates are bounded (Ahn et al., 2020, Nguyen et al., 2021). Besides this, Nguyen et al. (2021) have given a unified analysis for shuffling schemes other than RR. Safran and Shamir (2020) have provided a lower bound of $\Omega(\frac{1}{T^2} + \frac{n^2}{T^3})$ when f is a sum of n quadratics. In a more recent work, they have shown that RR does not significantly improve over SGD unless T is larger than κ in the worst case (Safran and Shamir, 2021). In this work, we show that *under interpolation RR can perform better than SGD when the number of samples (n) is less than the condition number, without requiring a large epoch, and without assuming component-wise convexity and G-Lipschitzness*. To our knowledge, this is the first work to show a gap in the convergence rates of RR and SGD under the over-parameterized setting.

2. Contributions

The convergence results of each algorithm are summarized in Table 1, and our main contributions are summarized as follows:

- Under interpolation (implied by the SGC), we derive the convergence rates of RR to be $\tilde{\mathcal{O}}(\kappa\sqrt{n}\sqrt{\rho+1+n})$ in the strongly-convex, $\tilde{\mathcal{O}}(\frac{L\sqrt{n}\sqrt{\rho+1+n}}{\epsilon})$ in the convex, and $\tilde{\mathcal{O}}(\frac{L\sqrt{n}\sqrt{\rho+1+n}}{\epsilon^2})$ in the non-convex cases, respectively. Comparing to RR, the rates of SGD under SGC are $\tilde{\mathcal{O}}(\kappa\rho)$ in the strongly-convex, $\tilde{\mathcal{O}}(\frac{L\rho+\rho-1}{\epsilon})$ in the convex, and $\tilde{\mathcal{O}}(\frac{L\rho}{\epsilon^2})$ in the non-convex cases, respectively (Vaswani et al., 2019). Hence, RR outperforms SGD when $\boldsymbol{\rho \gg n}$. Moreover, there is no requirement for a large epoch ($T \gtrsim \kappa$) as in previous works.

- While the choice of RR over SGD crucially depends on the relative magnitude of ρ and n, we show that RR consistently outperforms IG for both small and large ρ for all three objectives in Table 1.

- We conduct experiments on all the objectives considered by our theory. We verify that the condition $\boldsymbol{\rho \gg n}$ is important for RR to achieve a faster rate than SGD. Besides this, we demonstrate that RR converges faster than SGD in the early training phase and gains more advantages when n is small.

3. Convergence Results

In this section, we present the convergence results of RR for strongly-convex, convex, and non-convex objectives. Results for IG are given in Appendix D (see also Remark 2). Below, we use a constant step size η.

Theorem 1 (Strongly-Convex) *Suppose Assumptions 1,3 hold, assume f is μ-strongly convex, and choose a step size $\eta \leq \frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}$. Then, we have that*

$$\mathbb{E}[f(x_n^T) - f(x^*)] \leq \big(1 - \frac{1}{2}n\mu\eta\big)^T\big(f(x_0) - f^*\big) + \frac{4L^2\eta^2 n\sigma^2}{\mu}. \tag{6}$$

Moreover, under the SGC (Assumption 2), we obtain for $\eta = \frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}$ that

$$\mathbb{E}[f(x_n^T) - f(x^*)] \leq \big(1 - \frac{\sqrt{n}}{4\sqrt{2}\kappa\sqrt{\rho+1+n}}\big)^T\big(f(x_0) - f^*\big). \tag{7}$$

With a proper choice of η, we can translate (6) into a sample complexity of $\tilde{\mathcal{O}}(\kappa\sqrt{n}\sqrt{\rho+1+n} + \frac{\kappa\sqrt{n}\sigma}{\mu\sqrt{\epsilon}})$ (Corollary 1 in Appendix). Comparing this with $\tilde{\mathcal{O}}(\kappa\rho + \frac{\kappa\sqrt{\rho}\sigma}{\mu\sqrt{\epsilon}})$ for SGD (Vaswani et al., 2019), we see that RR performs better than SGD provided $n \ll \rho$. Setting $\sigma = 0$ reduces to the result shown in Table 1.

Remark 1 *Note that Theorem 1 holds true for any f that satisfies the PL condition (Lemma 1 in Appendix), which is more general than strong convexity.*

Theorem 2 (Convex) *Suppose Assumptions 1 and 2 hold, assume each f_i is convex, and set $\eta = \frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}$. Then, we obtain*

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \frac{2\sqrt{2}L\sqrt{\rho+1+n}\|x_0 - x^*\|^2}{T\sqrt{n}}, \tag{8}$$

where $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_0^t$.

The rate of RR in terms of sample complexity in the convex case is $\tilde{\mathcal{O}}(\frac{L\sqrt{n}\sqrt{\rho+1+n}}{\epsilon})$. This is better than $\tilde{\mathcal{O}}(\frac{Ln\sqrt{\rho}}{\epsilon})$ of IG when either $n \ll \rho$ or $n \gg \rho$, but only better than $\tilde{\mathcal{O}}(\frac{\rho L+\rho-1}{\epsilon})$ of SGD when $n \ll \rho$ (Vaswani et al., 2019).

Theorem 3 (Non-Convex) *Suppose Assumptions 1 and 3 hold, and choose a step size $\eta \leq \frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}$. Then, we have that*

$$\min_{t=1,\ldots T} \mathbb{E}[\|\nabla f(x_n^{t-1})\|^2] \leq \frac{4\big(f(x_0) - f^*\big)}{n\eta T} + 8L^2\eta^2 n\sigma^2. \tag{9}$$

Moreover, under the SGC (Assumption 2), we obtain for $\eta = \frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}$ that

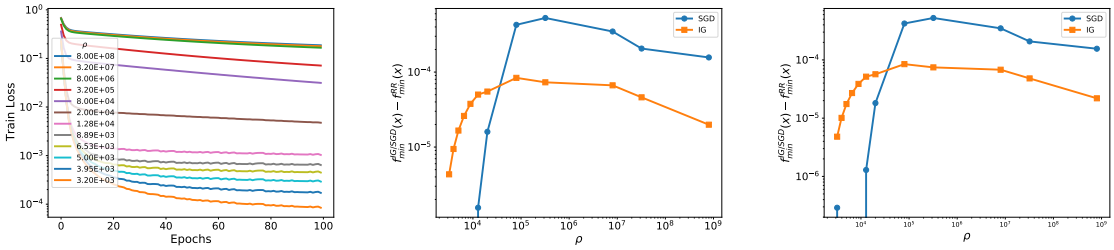
$$\min_{t=1,\ldots T} \mathbb{E}[\|\nabla f(x_n^{t-1})\|^2] \leq \frac{8L\sqrt{\rho+1+n}\big(f(x_0) - f^*\big)}{T\sqrt{n}}. \tag{10}$$

In the non-convex case, we can convert (9) into a sample complexity bound of $\tilde{\mathcal{O}}(\frac{L\sqrt{n}\sqrt{\rho+1+n}}{\epsilon^2} + \frac{L\sqrt{n}\sigma}{\epsilon^3})$ (Corollary 2 in Appendix). Setting $\sigma = 0$ and comparing this to the $\tilde{\mathcal{O}}(\frac{L\rho}{\epsilon^2})$ rate of SGD (Vaswani et al., 2019). We again observe that RR converges faster than SGD under SGC given $n \ll \rho$.

Remark 2 *The rates of IG are $\tilde{\mathcal{O}}(\kappa n\sqrt{\rho} + \frac{n\kappa\sigma}{\mu\sqrt{\epsilon}})$ and $\tilde{\mathcal{O}}(\frac{Ln\sqrt{\rho}}{\epsilon^2} + \frac{Ln\sigma}{\epsilon^3})$ for strongly-convex and non-convex cases respectively (Corollary 3 and 4 in the Appendix). These rates are not only worse than RR under SGC, but also worse in the case of $\sigma \neq 0$ by an extra factor of $\sqrt{n}$ shown in the 2nd term.*

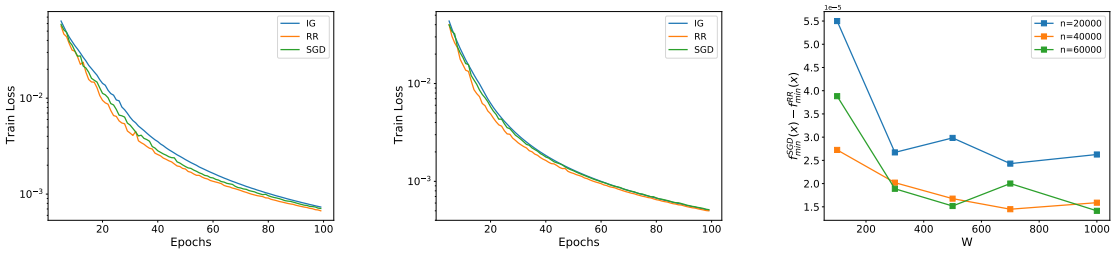
4. Experiments

We conduct experiments on a binary classification synthetic dataset and the MNIST image classification dataset (LeCun et al., 1998). To generate the synthetic dataset, we use the code provided by Loizou et al. (2021). Specifically, the data is generated to be linearly separable with margin $\tau > 0$. For training, we use a squared hinge loss with or without L_2 regularization to simulate the convex and the strongly-convex cases respectively. It has been shown that the squared hinge loss satisfies SGC with $\rho = \frac{n}{\tau^2}$ (Vaswani et al., 2019). Thus, varying τ, we investigate the convergence of SGD, RR, and IG as a function of the SGC constant ρ. The results are shown in Figure 1. In full agreement with our bounds, we observe that: (i) RR outperforms SGD when $\rho \gg n$, (ii) RR outperforms IG for all values of ρ, and (iii) the convergence (naturally) slows down for increasing ρ (which corresponds to decreasing margin τ). For image classification, we use a Softmax loss and one hidden-layer multilayer perceptrons (1-MLP) of different widths to simulate the non-convex setting. The results in Figure 2 show that RR can converge faster than SGD for a small T, and a small n favors RR while for large n there is less of an advantage to using RR.



(a) Train loss (b) Squared hinge loss (c) Squared hinge L_2 loss

Figure 1: Binary classification on a linearly separable dataset with $n = 800$. (a) We plot train loss of RR as a function of epochs using squared hinge loss for different values of ρ. We observe that the training is very slow when $\rho > \mathcal{O}(10^4)$, and speeds up when ρ approaches the order of magnitude of n or less. (b) We plot minimum train loss difference between RR and SGD or IG as a function of ρ. For a convex objective such as squared hinge loss, we observe that the performance gain of RR over SGD predominantly occurs when $\rho \gtrsim \mathcal{O}(10^4)$. For $\rho < \mathcal{O}(10^4)$, SGD achieves a lower train loss than RR. We also observe that RR consistently performs better than IG for all ρ values. (c) Similar observations are made as those of (b) for a strongly-convex objective: squared hinge loss with L_2 regularization.



(a) Train loss ($W = 100$) (b) Train loss ($W = 1000$) (c) Different ns

Figure 2: MNIST digits classification with a Softmax loss. (a) (b) We plot train loss as a function of epochs for 1-MLP of width (W) 100 and 1000 respectively. In both cases, we observe that RR converges faster than SGD in the initial training stage. (c) We plot minimum train loss difference between RR and SGD for $n = 20000, 40000$ and 60000 against different values of W. Despite RR outperforming SGD for all Ws and ns, the performance gap is more significant when $n = 20000$.

5. Conclusion

In this paper, we have derived convergence rates of Random Reshuffling under interpolation as implied by the SGC for strongly-convex, convex, and non-convex objectives. We do not have require a large epoch requirement and identify $\rho \ll n$ as being a crucial condition for Random Reshuffling to converge faster than SGD under all three assumptions. Our experimental results support these theory findings. In general, Random Reshuffling is favored under interpolation when the number of samples is small and/or the problem is ill-conditioned.

References

Kwangjun Ahn, Chulhee Yun, and Suvrit Sra. Sgd with shuffling: optimal rates without component convexity and large epoch requirements. *Advances in Neural Information Processing Systems*, 33:17526–17535, 2020.

Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

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

Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo A Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186(1):49–84, 2021.

Jeff Haochen and Suvrit Sra. Random shuffling beats sgd after finite epochs. In *International Conference on Machine Learning*, pages 2624–2633. PMLR, 2019.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR, 2021.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33:17309–17320, 2020.

Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. Sgd without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning*, pages 4703–4711. PMLR, 2019.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Lam M Nguyen, Quoc Tran-Dinh, Dzung T Phan, Phuong Ha Nguyen, and Marten Van Dijk. A unified convergence analysis for shuffling-type gradient methods. *The Journal of Machine Learning Research*, 22(1):9397–9440, 2021.

Shashank Rajput, Anant Gupta, and Dimitris Papailiopoulos. Closing the convergence gap of sgd without replacement. In *International Conference on Machine Learning*, pages 7964–7973. PMLR, 2020.

Itay Safran and Ohad Shamir. How good is sgd with random shuffling? In *Conference on Learning Theory*, pages 3250–3284. PMLR, 2020.

Itay Safran and Ohad Shamir. Random shuffling beats sgd only after many epochs on ill-conditioned problems. *Advances in Neural Information Processing Systems*, 34:15151–15161, 2021.

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

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR, 2019.

Organization of the Appendix We draw ideas from Nguyen et al. (2021) and Mishchenko et al. (2020) to construct our proofs. The high-level idea is to first bound the decrease in the objective value (see Lemma 5), then bound the progression term $\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2$ for IG and $\mathbb{E}[\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2]$ for RR respectively (see Lemma 6 and Lemma 7). Lemma 8 will be used in the proofs of convex objectives. All lemmas are given in Appendix B. All proofs of RR are given in Appendix C, which include the results of strongly-convex objective (Appendix C.1, C.4), convex objective (Appendix C.2), and non-convex objective (Appendix C.3, C.5). All proofs of IG are given in Appnedix D, which include the results of strongly-convex objective (Appendix D.1, D.4), convex objective (Appendix D.2), and non-convex objective (Appendix D.3, D.5).

Appendix A. Additional Convergence Results

Table 2: Number of gradient evaluations required by each algorithm to obtain an ϵ-accurate solution under Assumption 3. The ϵ-accurate solution is defined as $\|x - x^*\|^2 \leq \epsilon$, $f(x) - f(x^*) \leq \epsilon$, and $\|\nabla f(x)\| \leq \epsilon$ for μ-strongly convex, convex, and non-convex objectives respectively.

Citation	Algorithm	μ-Strongly Convex	Convex	Non-Convex
Vaswani et al. (2019)	SGD	$\kappa\rho + \frac{\sqrt{\rho}\kappa\sigma}{\mu\sqrt{\epsilon}}$	-	$\frac{L\rho}{\epsilon^2} + \frac{L\sigma^2}{\epsilon^4}$
This work	IG	$\kappa n\sqrt{\rho} + \frac{n\kappa\sigma}{\mu\sqrt{\epsilon}}$	-	$\frac{Ln\sqrt{\rho}}{\epsilon^2} + \frac{Ln\sigma}{\epsilon^3}$
	RR	$\kappa\sqrt{n}\sqrt{\rho+1+n} + \frac{\sqrt{n}\kappa\sigma}{\mu\sqrt{\epsilon}}$	-	$\frac{L\sqrt{n}\sqrt{\rho+1+n}}{\epsilon^2} + \frac{L\sqrt{n}\sigma}{\epsilon^3}$

Appendix B. Key Lemmas

A strongly-convex function f satisfies the PL condition stated in Lemma 1 (Karimi et al., 2016, Nesterov, 2003).

Lemma 1 (PL Condition) *A function f satisfies the PL condition if it satisfies the following inequality for some $\mu > 0$*

$$f(x) - f^* \leq \frac{1}{2\mu}\|\nabla f(x)\|^2 \quad \forall x \in dom(f), \tag{11}$$

where f^* is the optimal value of f.

The following lemma is a standard result for a μ-strongly convex function f (Nesterov, 2003).

Lemma 2 *For a μ-strongly convex function f, the following holds true*

$$\frac{\mu}{2}\|x - x^*\|^2 \leq f(x) - f(x^*) \quad \forall x \in dom(f), \tag{12}$$

where x^* is the unique minimizer of f.

Lemma 3 is a restatement of (Mishchenko et al., 2020, Lemma 1).

Lemma 3 *Let $X_1, ..., X_n$ be a given set of vectors in $\mathbb{R}^d$, denote their average to be $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and population variance to be $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \|X_i - \bar{X}\|^2$. Fix $k \in \{1, ..., n\}$, let $X_{\pi_1}, ..., X_{\pi_k}$ be sampled uniformly without replacement from $\{X_1, ..., X_n\}$ and $\bar{X}_\pi$ be their average. Then the following hold true*

$$\mathbb{E}[\bar{X}_\pi] = \bar{X} \qquad \mathbb{E}[\|\bar{X}_\pi - \bar{X}\|^2] = \frac{n-k}{k(n-1)}\sigma^2. \tag{13}$$

The remaining Lemmas 6-8 concern the update rules in Eqn (2). Recall that $x_0^{t+1} = x_n^t \ \forall t \geq 0$. Also recall that for RR the permutation vector π^t is chosen randomly at each epoch t while it is fixed for IG. The following lemma holds true for L-smooth and convex functions under SGC. We adapt the proof from (Nguyen et al., 2021, Lemma 7). It will be used in the proof of Lemma 8 below.

Lemma 4 *Suppose Assumptions 1,2 hold, and further assume each f_i is L_i-smooth and convex. For any permutation vector π^t, we have the following*

$$\sum_{j=0}^{n-1}\langle \nabla f(x_j^t; \pi_{j+1}^t), x^* - x_0^t\rangle \leq \frac{L}{2}\sum_{j=0}^{n-1}\|x_j^t - x_0^t\|^2 - \frac{1}{2L}\sum_{j=0}^{n-1}\|\nabla f(x_j^t; \pi_{j+1}^t)\|^2 - n(f(x_0^t) - f(x^*)). \tag{14}$$

Proof By (Nguyen et al., 2021, Lemma 7),

$$\sum_{j=0}^{n-1}\langle \nabla f(x_j^t; \pi_{j+1}^t), x^* - x_0^t\rangle \leq \frac{L}{2}\sum_{j=0}^{n-1}\|x_j^t - x_0^t\|^2 - \frac{1}{2L}\sum_{j=0}^{n-1}\|\nabla f(x_j^t; \pi_{j+1}^t) - \nabla f(x^*; \pi_{j+1}^t)\|^2 - n(f(x_0^t) - f(x^*)). \tag{15}$$

By SGC, which implies $\nabla f(x^*; \pi_{j+1}^t) = 0$ for any π^t, the desired result follows. ■

Lemma 5 is proved in (Nguyen et al., 2021, Lemma 8). Here we take the step size η to be constant for all iterations and epochs.

Lemma 5 *Suppose Assumption 1 holds. Given a shuffling scheme $\{\pi^t\}_t$ and a constant step size η such that $\eta \leq \frac{1}{nL}$, we have the following:*

$$f(x_0^{t+1}) \leq f(x_0^t) - \frac{n\eta}{2}\|\nabla f(x_0^t)\|^2 + \frac{L^2\eta}{2}\sum_{i=0}^{n}\|x_i^t - x_0^t\|^2. \tag{16}$$

Taking total expectation over the randomness

$$\mathbb{E}[f(x_0^{t+1})] \leq \mathbb{E}[f(x_0^t)] - \frac{n\eta}{2}\mathbb{E}[\|\nabla f(x_0^t)\|^2] + \frac{L^2\eta}{2}\sum_{i=0}^{n}\mathbb{E}[\|x_i^t - x_0^t\|^2]. \tag{17}$$

Lemma 6 will be combined with Lemma 5 in the proofs of IG.

Lemma 6 *Suppose Assumptions 1, 3 hold. For a step size $\eta \leq \frac{1}{\sqrt{2}nL}$, it holds for IG that*

$$\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2 \leq 2\eta^2 n^3 \rho \|\nabla f(x_0^t)\|^2 + 2\eta^2 n^3 \sigma^2. \tag{18}$$

Proof By the update rule

$$\begin{aligned}
\|x_i^t - x_0^t\|^2 &= \eta^2 \|\sum_{j=0}^{i-1} \nabla f(x_j^t; j+1)\|^2 \\
&= \eta^2 \|\sum_{j=0}^{i-1} \nabla f(x_j^t; j+1) - \nabla f(x_0^t; j+1) + \nabla f(x_0^t; j+1)\|^2 \\
&\leq 2\eta^2 \|\sum_{j=0}^{i-1} \nabla f(x_j^t; j+1) - \nabla f(x_0^t; j+1)\|^2 + 2\eta^2 \|\sum_{j=0}^{i-1} \nabla f(x_0^t; j+1)\|^2 \\
&\leq 2\eta^2 i \sum_{j=0}^{i-1} \|\nabla f(x_j^t; j+1) - \nabla f(x_0^t; j+1)\|^2 + 2\eta^2 i \sum_{j=0}^{i-1} \|\nabla f(x_0^t; j+1)\|^2 \\
&\leq 2\eta^2 L^2 i \sum_{j=0}^{i-1} \|x_j^t - x_0^t\|^2 + 2\eta^2 i \sum_{j=0}^{i-1} \|\nabla f(x_0^t; j+1)\|^2.
\end{aligned} \tag{19}$$

Summing over $i = 0, ..., n-1$ gives

$$\begin{aligned}
\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2 &\leq 2\eta^2 L^2 \sum_{j=0}^{i-1} \|x_j^t - x_0^t\|^2 \sum_{i=0}^{n-1} i + 2\eta^2 \sum_{j=0}^{n-1} \|\nabla f(x_0^t; j+1)\|^2 \sum_{i=0}^{n-1} i \\
&\leq \eta^2 L^2 n^2 \sum_{j=0}^{n-1} \|x_j^t - x_0^t\|^2 + \eta^2 n^3 \frac{1}{n} \sum_{j=0}^{n-1} \|\nabla f(x_0^t; j+1)\|^2 \\
&\leq \eta^2 L^2 n^2 \sum_{j=0}^{n-1} \|x_j^t - x_0^t\|^2 + \eta^2 n^3 (\rho \|\nabla f(x_0^t)\|^2 + \sigma^2) \\
&= \eta^2 L^2 n^2 \sum_{j=0}^{n-1} \|x_j^t - x_0^t\|^2 + \eta^2 n^3 \rho \|\nabla f(x_0^t)\|^2 + \eta^2 n^3 \sigma^2.
\end{aligned} \tag{20}$$

Finally, choosing $\eta \leq \frac{1}{\sqrt{2}nL}$ leads to

$$\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2 \leq 2\eta^2 n^3 \rho \|\nabla f(x_0^t)\|^2 + 2\eta^2 n^3 \sigma^2. \tag{21}$$

■

Lemma 7 below will be combined with Lemma 5 in the proofs of RR.

Lemma 7 *Suppose Assumption 1, 3 hold. For a step size $\eta \leq \frac{1}{\sqrt{3}nL}$, the following holds for RR*

$$\mathbb{E}\Big[\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2\Big] \leq 4n^2\eta^2(\rho+1+n)\mathbb{E}[\|\nabla f(x_0^t)\|^2] + 4\eta^2 n^2\sigma^2. \tag{22}$$

Proof By the update rule

$$\begin{aligned}
\|x_i^t - x_0^t\|^2 &= \eta^2\|\sum_{j=0}^{i-1}\nabla f(x_j^t;\pi_{j+1}^t)\|^2 \\
&= \eta^2\|\sum_{j=0}^{i-1}\nabla f(x_j^t;\pi_{j+1}^t) - \nabla f(x_0^t;\pi_{j+1}^t) + \nabla f(x_0^t;\pi_{j+1}^t) - \nabla f(x_0^t) + \nabla f(x_0^t)\|^2 \\
&\leq 3\eta^2\|\sum_{j=0}^{i-1}\nabla f(x_j^t;\pi_{j+1}) - \nabla f(x_0^t;\pi_{j+1}^t)\|^2 + 3\eta^2\|\sum_{j=0}^{i-1}\nabla f(x_0^t;\pi^t j+1) \\
&\quad - \nabla f(x_0^t)\|^2 + 3\eta^2\|\sum_{j=0}^{i-1}\nabla f(x_0^t)\|^2 \\
&\leq 3\eta^2 i\sum_{j=0}^{i-1}\|\nabla f(x_j^t;\pi_{j+1}) - \nabla f(x_0^t;\pi_{j+1}^t)\|^2 + 3\eta^2 i^2\|\frac{1}{i}\sum_{j=0}^{i-1}\nabla f(x_0^t;\pi^t j+1) \\
&\quad - \nabla f(x_0^t)\|^2 + 3\eta^2 i^2\|\nabla f(x_0^t)\|^2 \\
&\leq 3\eta^2 iL^2\sum_{j=0}^{i-1}\|x_j^t - x_0^t\|^2 + 3\eta^2 i^2\|\frac{1}{i}\sum_{j=0}^{i-1}\nabla f(x_0^t;\pi^t j+1) - \nabla f(x_0^t)\|^2 \\
&\quad + 3\eta^2 i^2\|\nabla f(x_0^t)\|^2.
\end{aligned} \tag{23}$$

Let σ^t be a sigma algebra on the iterates $\{x_0^1, ..., x_0^t\}$. We take conditional expectation w.r.t σ^t and apply Lemma 3 to find that

$$\begin{aligned}
\mathbb{E}\Big[\|\frac{1}{i}\sum_{j=0}^{i-1}\nabla f(x_0^t;\pi_{j+1}^t) - \nabla f(x_0^t)\|^2|\sigma^t\Big] &= \frac{n-i}{i(n-1)}\frac{1}{n}\sum_{j=0}^{n-1}\|\nabla f(x_0^t;j+1) - \nabla f(x_0^t)\|^2 \\
&\leq \frac{n-i}{i(n-1)}\frac{1}{n}\sum_{j=0}^{n-1}[2\|\nabla f(x_0^t;j+1)\|^2 + 2\|\nabla f(x_0^t)\|^2].
\end{aligned} \tag{24}$$

Take conditional expectation of (23) and substitute (24) back

$$\begin{aligned}\mathbb{E}\Big[\|x_i^t - x_0^t\||\sigma^t\Big] &\leq 3\eta^2 iL^2 \sum_{j=0}^{n-1} \mathbb{E}[\|x_j^t - x_0^t\||\sigma^t] + 6\eta^2 \frac{i(n-i)}{(n-1)} \frac{1}{n} \sum_{j=0}^{n-1} \|\nabla f(x_0^t; j+1)\|^2 + \\ &\quad 6\eta^2 \frac{i(n-i)}{n-1} \|\nabla f(x_0^t)\|^2 + 3\eta^2 i^2 \|\nabla f(x_0^t)\|^2 \\ &\leq 3\eta^2 iL^2 \sum_{j=0}^{n-1} \mathbb{E}[\|x_j^t - x_0^t\||\sigma^t] + 6\eta^2 \frac{i(n-i)}{n-1} (\rho \|\nabla f(x_0^t)\|^2 + \sigma^2) + \\ &\quad 6\eta^2 \frac{i(n-i)}{n-1} \|\nabla f(x_0^t)\|^2 + 3\eta^2 i^2 \|\nabla f(x_0^t)\|^2. \end{aligned} \tag{25}$$

Next, take total expectation and sum over $i = 0, ..., n-1$:

$$\begin{aligned}\mathbb{E}\Big[\sum_{i=0}^{n-1} \|x_i^t - x_0^t\|^2\Big] &\leq 3\eta^2 L^2 \mathbb{E}[\sum_{j=0}^{n-1} \|x_j^t - x_0^t\|^2] \sum_{i=0}^{n-1} i + 6\eta^2 \frac{1}{n-1} (\rho \mathbb{E}[\|\nabla f(x_0^t)\|^2] + \sigma^2) \sum_{i=0}^{n-1} i(n-i) + \\ &\quad 6\eta^2 \frac{1}{n-1} \mathbb{E}[\|\nabla f(x_0^t)\|^2] \sum_{i=0}^{n-1} i(n-i) + 3\eta^2 \mathbb{E}[\|\nabla f(x_0^t)\|^2] \sum_{i=0}^{n-1} i^2 \\ &\leq \frac{3}{2} \eta^2 L^2 n^2 \mathbb{E}[\sum_{j=0}^{n-1} \|x_j^t - x_0^t\|^2] + 2\eta^2 n^2 \rho \mathbb{E}[\|\nabla f(x_0^t)\|^2] + 2\eta^2 n^2 \sigma^2 + \\ &\quad 2\eta^2 n^2 \mathbb{E}[\|\nabla f(x_0^t)\|^2] + 2\eta^2 n^3 \mathbb{E}[\|\nabla f(x_0^t)\|^2] \qquad (26) \\ &= \frac{3}{2} \eta^2 L^2 n^2 \mathbb{E}[\sum_{j=0}^{n-1} \|x_j^t - x_0^t\|^2] + 2\eta^2 n^2 (\rho + 1 + n) \mathbb{E}[\|\nabla f(x_0^t)\|^2] + 2\eta^2 n^2 \sigma^2. \qquad (27)\end{aligned}$$

In (26), we have used $\sum_{i=0}^{n-1} i \leq \frac{n^2}{2}$, $\sum_{i=0}^{n-1} i^2 \leq \frac{n^3}{3}$, and $\sum_{i=0}^{n-1} i(n-i) \leq \frac{n^2(n-1)}{3}$. Choosing $\eta \leq \frac{1}{\sqrt{3}Ln}$, we have

$$\mathbb{E}[\sum_{i=0}^{n-1} \|x_i^t - x_0^t\|^2] \leq 4\eta^2 n^2 (\rho + 1 + n) \mathbb{E}[\|\nabla f(x_0^t)\|^2] + 4\eta^2 n^2 \sigma^2. \tag{28}$$

■

Lemma 8 will be used in the proofs of convex objectives.

Lemma 8 *Suppose Assumptions 1, 2 hold. Assume each f_i is convex, choose a learning rate $\eta \leq \frac{1}{nL}$, then we have the following*

$$\|x_n^t - x^*\|^2 \leq \|x_0^t - x^*\|^2 + \eta L \sum_{j=0}^{n-1} \|x_j^t - x_0^t\|^2 - 2\eta n [f(x_0^t) - f(x^*)]. \tag{29}$$

Proof

$$\|x_n^t - x^*\|^2 = \|x_0^t - x^*\|^2 + 2\eta \sum_{j=0}^{n-1} \langle \nabla f(x_j^t; \pi_{j+1}^t), x^* - x_0^t \rangle + \eta^2 \|\sum_{j=0}^{n-1} \nabla f(x_j^t; \pi_{j+1}^t)\|^2$$
$$\leq \|x_0^t - x^*\|^2 + 2\eta \sum_{j=0}^{n-1} \langle \nabla f(x_j^t; \pi_{j+1}^t), x^* - x_0^t \rangle + \eta^2 n \sum_{j=0}^{n-1} \|\nabla f(x_j^t; \pi_{j+1}^t)\|^2 \quad (30)$$

By Lemma 4, we have

$$\sum_{j=0}^{n-1} \langle \nabla f(x_j^t; \pi_{j+1}^t), x^* - x_0^t \rangle \leq \frac{L}{2} \sum_{j=0}^{n-1} \|x_j^t - x_0^t\|^2 -$$
$$\frac{1}{2L} \sum_{j=0}^{n-1} \|\nabla f(x_j^t; \pi_{j+1}^t)\|^2 - n(f(x_0^t) - f(x^*)). \quad (31)$$

Substitute this back in

$$\|x_n^t - x^*\|^2 \leq \|x_0^t - x^*\|^2 + \eta L \sum_{j=0}^{n-1} \|x_j^t - x_0^t\|^2 - \frac{\eta}{L} \sum_{j=0}^{n-1} \|\nabla f(x_j^t; \pi_{j+1}^t)\|^2 -$$
$$2\eta n(f(x_0^t) - f(x^*)) + \eta^2 n \sum_{j=0}^{n-1} \|\nabla f(x_j^t; \pi_{j+1}^t)\|^2 \quad (32)$$
$$= \|x_0^t - x^*\|^2 + \eta L \sum_{j=0}^{n-1} \|x_j^t - x_0^t\|^2 - 2\eta n(f(x_0^t) - f(x^*)) +$$
$$(\eta^2 n - \frac{\eta}{L}) \sum_{j=0}^{n-1} \|\nabla f(x_j^t; \pi_{j+1}^t)\|^2. \quad (33)$$

Choose $\eta \leq \frac{1}{nL}$, we have $\eta^2 n - \frac{\eta}{L} \leq 0$, then

$$\|x_n^t - x^*\|^2 \leq \|x_0^t - x^*\|^2 + \eta L \sum_{j=0}^{n-1} \|x_j^t - x_0^t\|^2 - 2\eta n[f(x_0^t) - f(x^*)]. \quad (34)$$

■

Appendix C. Convergence Proofs of RR

C.1. Proof of Theorem 1

Apply Lemma 5 and Lemma 7

$$\mathbb{E}[f(x_0^{t+1})] \leq \mathbb{E}[f(x_0^t)] - \frac{n\eta}{2} \mathbb{E}[\|\nabla f(x_0^t)\|^2] + \frac{L^2 \eta}{2} [4\eta^2 n^2 (\rho + 1 + n) \mathbb{E}[\|\nabla f(x_0^t)\|^2] + 4\eta^2 n^2 \sigma^2]$$
$$= \mathbb{E}[f(x_0^t)] - \frac{n\eta}{2} [1 - 4L^2 \eta^2 n(\rho + 1 + n)] \mathbb{E}[\|\nabla f(x_0^t)\|^2] + 2L^2 \eta^3 n^2 \sigma^2. \quad (35)$$

Further use the PL condition in Lemma 1 to arrive at

$$\mathbb{E}[f(x_0^{t+1}) - f(x^*)] \leq \mathbb{E}[f(x_0^t) - f(x^*)] - n\eta\mu[1 - 4L^2\eta^2 n(\rho + 1 + n)]\mathbb{E}[f(x_0^t) - f(x^*)] + 2L^2\eta^3 n^2\sigma^2. \tag{36}$$

Choosing $\eta \leq \frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}$ gives

$$\mathbb{E}[f(x_0^{t+1}) - f(x^*)] \leq (1 - \frac{1}{2}n\mu\eta)\mathbb{E}[f(x_0^t) - f(x^*)] + 2L^2\eta^3 n^2\sigma^2. \tag{37}$$

Solving (37) recursively for $t = 1, ..., T$ gives

$$\mathbb{E}[f(x_0^{T+1}) - f(x^*)] \leq (1 - \frac{1}{2}n\mu\eta)^T(f(x_0^1) - f(x^*)) + 2L^2\eta^3 n^2\sigma^2 \sum_{j=0}^{T}(1 - \frac{1}{2}n\mu\eta)^j. \tag{38}$$

This implies (recalling that $x_0^1 = x_0$):

$$\begin{aligned}\mathbb{E}[f(x_n^T) - f(x^*)] &\leq (1 - \frac{1}{2}n\mu\eta)^T(f(x_0^1) - f(x^*)) + 2L^2\eta^3 n^2\sigma^2 \sum_{j=0}^{\infty}(1 - \frac{1}{2}n\mu\eta)^j \\ &\leq (1 - \frac{1}{2}n\mu\eta)^T(f(x_0) - f(x^*)) + 2L^2\eta^3 n^2\sigma^2(\frac{2}{n\mu\eta}) \\ &= (1 - \frac{1}{2}n\mu\eta)^T(f(x_0) - f(x^*)) + \frac{4L^2\eta^2 n\sigma^2}{\mu}. \end{aligned} \tag{39}$$

Note that for (39) to hold, we require $\eta \leq \min\{\frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}, \frac{1}{\sqrt{3}Ln}\}$ because of Lemma 7. This condition becomes $\eta \leq \frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}$ as $\frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}} \leq \frac{1}{\sqrt{3}Ln}$. Under SGC, substitute $\sigma = 0$ with $\eta = \frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}$ to obtain the desired result.

C.2. Proof of Theorem 2

Start with Lemma 8

$$\|x_n^t - x^*\|^2 \leq \|x_0^t - x^*\|^2 + \eta L \sum_{j=0}^{n-1}\|x_j^t - x_0^t\|^2 - 2\eta n[f(x_0^t) - f(x^*)]. \tag{40}$$

Take expectation and apply Lemma 7 with $\sigma = 0$

$$\begin{aligned}\mathbb{E}[\|x_n^t - x^*\|^2] &\leq \mathbb{E}[\|x_0^t - x^*\|^2] + \eta L(4\eta^2 n^2(\rho + 1 + n)\mathbb{E}[\|\nabla f(x_0^t)\|^2]) - \\ &\quad 2\eta n\mathbb{E}[f(x_0^t) - f(x^*)] \\ &\leq \mathbb{E}[\|x_0^t - x^*\|^2] - 2\eta n[1 - 4\eta^2 n L^2(\rho + 1 + n)]\mathbb{E}[f(x_0^t) - f(x^*)]. \end{aligned} \tag{41}$$

Choose $\eta \leq \frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}$:

$$\eta n\mathbb{E}[f(x_0^t) - f(x)] \leq \mathbb{E}[\|x_0^t - x^*\|^2] - \mathbb{E}[\|x_0^{t+1} - x^*\|^2]. \tag{42}$$

Sum $t = 1, ..., T$, and divide $\eta n T$ on both sides

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[f(x_0^t) - f(x^*)] \leq \frac{\|x_0 - x^*\|^2}{T\eta n}. \tag{43}$$

Define $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_0^t$. By convexity of f and Jensen's inequality:

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \frac{1}{\eta n T}\|x_0 - x^*\|^2. \tag{44}$$

Substitute $\eta = \frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}$ to obtain the result for SGC.

C.3. Proof of Theorem 3

From (35) in the Proof of Theorem 1

$$\mathbb{E}[f(x_0^{t+1})] \leq \mathbb{E}[f(x_0^t)] - \frac{n\eta}{2}[1 - 4L^2\eta^2 n(\rho + 1 + n)]\mathbb{E}[\|\nabla f(x_0^t)\|^2] + 2L^2\eta^2 n\sigma^2. \tag{45}$$

Choose $\eta \leq \frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}$:

$$\mathbb{E}[f(x_0^{t+1})] \leq \mathbb{E}[f(x_0^t)] - \frac{n\eta}{4}\mathbb{E}[\|\nabla f(x_0^t)\|^2] + 2L^2\eta^3 n^2\sigma^2. \tag{46}$$

Rearrange and sum over $t = 1, ..., T$

$$\frac{n\eta}{4}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x_0^t)\|^2 \leq \mathbb{E}[f(x_0^1)] + 2L^2\eta^3 n^2\sigma^2 T. \tag{47}$$

Divide $\frac{n\eta T}{4}$ on both sides:

$$\min_{t=1,...T}\mathbb{E}[\|\nabla f(x_0^t)\|^2] \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(x_0^t)\|^2] \leq \frac{4\mathbb{E}[f(x_0) - f(x^*)]}{n\eta T} + 8L^2\eta^2 n\sigma^2. \tag{48}$$

Substitute $\eta = \frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}$ and we obtain the result for SGC.

C.4. Sample Complexity for Strongly-Convex Objective

Corollary 1 *Suppose Assumption 1, 3 hold. Assume f is μ-strongly convex, choose $\eta = \min\{\frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}, \frac{2}{n\mu T}\log(\frac{(f(x_0)-f(x^*))\mu^3 T^2 n}{L^2\sigma^2})\}$. Then we have*

$$\mathbb{E}[f(x_n^T) - f(x^*)] = \tilde{\mathcal{O}}\Big(\exp\big(-\frac{T\sqrt{n}}{\kappa\sqrt{\rho+1+n}}\big)\big(f(x_0) - f(x^*)\big) + \frac{\kappa^2\sigma^2}{n\mu T^2}\Big). \tag{49}$$

Proof For the case $\eta = \frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}} \leq \frac{2}{n\mu T}\log(\frac{(f(x_0)-f(x^*))\mu^3T^2n}{L^2\sigma^2})$, substitute this η into (39)

$$\begin{aligned}\mathbb{E}[f(x_n^T)-f(x^*)] &\leq (1-\frac{\sqrt{n}}{4\sqrt{2}\kappa\sqrt{\rho+1+n)}})^T(f(x_0)-f(x^*))+\frac{\sigma^2}{2\mu(\rho+1+n)}\\ &\leq \exp(-\frac{T\sqrt{n}}{4\sqrt{2}\kappa\sqrt{\rho+1+n}})(f(x_0)-f(x^*))+\frac{16L^2\sigma^2}{n\mu^3T^2}\log^2(\frac{(f(x_0)-f(x^*)\mu^3T^2n}{L^2\sigma^2})\\ &= \tilde{\mathcal{O}}\Big(\exp\big(-\frac{T\sqrt{n}}{\kappa\sqrt{\rho+1+n}}\big)\big)\big(f(x_0)-f(x^*)\big)+\frac{\kappa^2\sigma^2}{n\mu T^2}\Big). \qquad (50)\end{aligned}$$

For the case $\eta = \frac{2}{n\mu T}\log(\frac{(f(x_0)-f(x^*))\mu^3T^2n}{L^2\sigma^2}) \leq \frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}$, we have

$$\begin{aligned}\mathbb{E}[f(x_n^T)-f(x^*)] \leq \exp(-\frac{n\mu T}{2}\frac{2}{n\mu T}\log(\frac{((f(x_0)-f(x^*)))\mu^3T^2n}{L^2\sigma^2})(f(x_0)-f(x^*))+\\ \frac{4L^2n\sigma^2}{\mu}\frac{4}{n^2\mu^2T^2}\log^2(\frac{(f(x_0)-f(x^*))\mu^3T^2n}{L^2\sigma^2})\\ = \frac{L^2\sigma^2}{\mu^3T^2n}+\frac{16L^2\sigma^2}{\mu^3T^2n}\log^2(\frac{(f(x_0)-f(x^*))\mu^3T^2n}{L^2\sigma^2})\\ = \tilde{\mathcal{O}}(\frac{\kappa^2\sigma^2}{n\mu T^2}). \qquad (51)\end{aligned}$$

Combining (50) and (51), we obtain

$$\mathbb{E}[f(x_n^T)-f(x^*)] = \tilde{\mathcal{O}}\Big(\exp\big(-\frac{T\sqrt{n}}{\kappa\sqrt{\rho+1+n}}\big)\Delta_0+\frac{\kappa^2\sigma^2}{n\mu T^2}\Big), \qquad (52)$$

where $\Delta_0 = f(x_0)-f(x^*)$. Apply Lemma 2 to (52)

$$\mathbb{E}[\|x_n^T-x^*\|^2] = \tilde{\mathcal{O}}\Big(\exp\big(-\frac{T\sqrt{n}}{\kappa\sqrt{\rho+1+n}}\big)\frac{\Delta_0}{\mu}+\frac{\kappa^2\sigma^2}{n\mu^2T^2}\Big). \qquad (53)$$

This translates to a sample complexity of $\tilde{\mathcal{O}}(\kappa\sqrt{n}\sqrt{\rho+1+n}+\frac{\kappa\sqrt{n}\sigma}{\mu\sqrt{\epsilon}})$ where logarithmic factors are ignored. ■

C.5. Sample Complexity of Non-Convex Objective

Corollary 2 *Suppose Assumption 1,3 hold, assume f is Non-Convex, then the sample complexity for f to reach an ϵ-accurate solution ($\|\nabla f(x)\| \leq \epsilon$) is*

$$\tilde{\mathcal{O}}(\frac{L\sqrt{n}\sqrt{\rho+1+n}}{\epsilon^2}+\frac{L\sqrt{n}\sigma}{\epsilon^3}). \qquad (54)$$

Proof Start from (48) in the Proof of Theorem 3

$$\min_{t=1,\dots T}\mathbb{E}[\|\nabla f(x_0^t)\|^2] \leq \frac{4\mathbb{E}[f(x_0)-f(x^*)]}{n\eta T}+8L^2\eta^2n\sigma^2. \qquad (55)$$

We first require $8L^2\eta^2 n\sigma^2 \leq \frac{\epsilon^2}{2}$, this implies

$$\eta \leq \frac{\epsilon}{4L\sqrt{n}\sigma}. \tag{56}$$

Note that for (55) to hold, we require $\eta \leq \frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}$, combine this with (56), we have

$$\eta \leq \min\{\frac{1}{2\sqrt{2}L\sqrt{n(\rho+1+n)}}, \frac{\epsilon}{4L\sqrt{n}\sigma}\}. \tag{57}$$

Note that $\frac{4\mathbb{E}[f(x_0)-f(x^*)]}{n\eta T} \leq \frac{\epsilon^2}{2} \Rightarrow \frac{8[f(x_0^t-f(x^*))]}{\eta\epsilon^2} \leq nT$, this implies

$$\begin{aligned} nT \geq \frac{8[f(x_0^t - f(x^*))]}{\eta\epsilon^2} &= \frac{8[f(x_0^t) - f(x^*))]}{\epsilon^2}\max\{2\sqrt{2}L\sqrt{n(\rho+1+n)}, \frac{4L\sqrt{n}\sigma}{\epsilon}\} \\ &= \frac{32[f(x_0^t) - f(x^*)]L\sqrt{n}}{\epsilon^2}\max\{\sqrt{\rho+1+n}, \frac{\sigma}{\epsilon}\} \\ &= \tilde{\mathcal{O}}(\frac{L\sqrt{n}\sqrt{\rho+1+n}}{\epsilon^2} + \frac{L\sqrt{n}\sigma}{\epsilon^3}). \end{aligned} \tag{58}$$

■

Appendix D. Convergence Results and Proofs of IG

D.1. Convergence of Strongly-Convex Objective

Theorem 4 *Suppose Assumptions 1,3 hold. Assume f is μ-strongly convex and choose a step size $\eta \leq \frac{1}{2nL\sqrt{\rho}}$. Then,*

$$f(x_n^T) - f(x^*) \leq (1 - \frac{1}{2}n\mu\eta)^T(f(x_0) - f(x^*)) + \frac{2L^2\eta^2n^2\sigma^2}{\mu}. \tag{59}$$

Under SGC (Assumption 2), set $\eta = \frac{1}{2nL\sqrt{\rho}}$ to obtain

$$f(x_n^T) - f(x^*) \leq (1 - \frac{1}{4\kappa\sqrt{\rho}})^T(f(x_0) - f(x^*)). \tag{60}$$

Proof Apply Lemma 5 and Lemma 6

$$\begin{aligned} f(x_0^{t+1}) &\leq f(x_0^t) - \frac{n\eta}{2}\|\nabla f(x_0^t)\|^2 + \frac{L^2\eta}{2}(2\eta^2n^3\rho\|\nabla f(x_0^t)\|^2 + 2\eta^2n^3\sigma^2) \\ &= f(x_0^t) - \frac{n\eta}{2}(1 - 2L^2\eta^2n^2\rho)\|\nabla f(x_0^t)\|^2 + L^2\eta^3n^3\sigma^2. \end{aligned} \tag{61}$$

Choose $\eta \leq \frac{1}{2nL\sqrt{\rho}}$ and apply Lemma 1

$$f(x_0^{t+1}) - f(x^*) \leq (1 - \frac{1}{2}n\mu\eta)(f(x_0^t) - f(x^*)) + L^2\eta^3n^3\sigma^2. \tag{62}$$

Solve (62) recursively

$$f(x_0^{T+1}) - f(x^*) \leq (1 - \frac{1}{2}n\mu\eta)^T(f(x_0^1) - f(x^*)) + L^2\eta^3n^3\sigma^2\sum_{j=0}^{T}(1 - \frac{1}{2}n\mu\eta)^j. \tag{63}$$

This implies

$$\begin{aligned} f(x_n^T) - f(x^*) &\leq (1 - \frac{1}{2}n\mu\eta)^T(f(x_0) - f(x^*)) + L^2\eta^3n^3\sigma^2\sum_{j=0}^{\infty}(1 - \frac{1}{2}n\mu\eta)^j \\ &\leq (1 - \frac{1}{2}n\mu\eta)^T(f(x_0) - f(x^*)) + L^2\eta^3n^3\sigma^2(\frac{2}{n\mu\eta}) \\ &= (1 - \frac{1}{2}n\mu\eta)^T(f(x_0) - f(x^*)) + \frac{2L^2\eta^2n^2\sigma^2}{\mu}. \end{aligned} \tag{64}$$

Note for (64) to hold, we require $\eta \leq min\{\frac{1}{2nL\sqrt{\rho}}, \frac{1}{\sqrt{2}nL}\}$ as we used Lemma 6. This reduces to $\eta \leq \frac{1}{2nL\sqrt{\rho}}$ provided $\rho \geq 1$. Under SGC, set $\sigma = 0$ and substitute $\eta = \frac{1}{2nL\sqrt{\rho}}$, we obtain the desired result. ■

D.2. Convergence of Convex Objective

Theorem 5 *Suppose Assumptions 1,2 hold. Assume each f_i is convex. Set $\eta = \frac{1}{2Ln\sqrt{\rho}}$. Then*

$$f(\bar{x}_T) - f(x^*) \leq \frac{2L\sqrt{\rho}\|x_0 - x^*\|^2}{T}, \tag{65}$$

where $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_0^t$.

Proof Start with Lemma 8

$$\|x_n^t - x^*\|^2 \leq \|x_0^t - x^*\|^2 + \eta L\sum_{j=0}^{n-1}\|x_j^t - x_0^t\|^2 - 2\eta n[f(x_0^t) - f(x^*)]. \tag{66}$$

Use the result of Lemma 6 with $\sigma = 0$

$$\begin{aligned} \|x_n^t - x^*\|^2 &\leq \|x_0^t - x^*\|^2 + \eta L(2\eta^2n^3\rho\|\nabla f(x_0^t)\|^2) - 2\eta n(f(x_0^t - f(x^*)) \\ &= \|x_0^t - x^*\|^2 + 2n^3\eta^3L\rho\|\nabla f(x_0^t)\|^2 - 2\eta n(f(x_0^t) - f(x^*)) \\ &\leq \|x_0^t - x^*\|^2 + 2n^3\eta^3L\rho(2L(f(x_0^t) - f(x^*))) - 2\eta n(f(x_0^t) - f(x^*)) + \\ &\quad 2L\eta^3n^3\sigma^2 \\ &\leq \|x_0^t - x^*\|^2 - 2\eta n(1 - 2L^2n^2\eta^2\rho)(f(x_0^t) - f(x^*)). \end{aligned} \tag{67}$$

Choose $\eta \leq \frac{1}{2Ln\sqrt{\rho}}$

$$\eta n(f(x_0^t) - f(x)) \leq \|x_0^t - x^*\| + \|x_0^{t+1} - x^*\|^2. \tag{68}$$

Define $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_0^t$. By convexity of f

$$f(\bar{x}_T) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{T\eta n}, \tag{69}$$

where $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_0^t$. Note for (69) to hold, we require $\eta \leq \min\{\frac{1}{\sqrt{2}nL}, \frac{1}{2Ln\sqrt{\rho}}\}$ as Lemma 6 is used. This reduces to $\eta \leq \frac{1}{2Ln\sqrt{\rho}}$ provided $\rho \geq 1$. Substitute $\eta = \frac{1}{2Ln\sqrt{\rho}}$ to obtain the desired result. ■

D.3. Convergence of Non-Convex Objective

Theorem 6 *Suppose Assumptions 1,3 hold, choose a step size $\eta \leq \frac{1}{2nL\sqrt{\rho}}$, we have*

$$\min_{t=1,\ldots T} \|\nabla f(x_n^{t-1})\|^2 \leq \frac{4(f(x_0) - f(x^*)))}{n\eta T} + 4L^2\eta^2 n^2 \sigma^2. \tag{70}$$

Under SGC (Assumption 2), set $\eta = \frac{1}{2nL\sqrt{\rho}}$ to obtain

$$\min_{t=1,\ldots T} \|\nabla f(x_n^{t-1})\|^2 \leq \frac{8\sqrt{2}L\sqrt{\rho}(f(x_0) - f(x^*))}{T}. \tag{71}$$

Proof From (61) in the Proof of Theorem 4

$$f(x_0^{t+1}) \leq f(x_0^t) - \frac{n\eta}{2}(1 - 2L^2\eta^2 n^2 \rho)\|\nabla f(x_0^t)\|^2 + L^2\eta^3 n^3 \sigma^2. \tag{72}$$

Choose $\eta \leq \frac{1}{2nL\sqrt{\rho}}$

$$f(x_0^{t+1}) \leq f(x_0^t) - \frac{n\eta}{4}\|\nabla f(x_0^t)\|^2 + L^2\eta^3 n^3 \sigma^2. \tag{73}$$

Rearrange and sum over $t = 1, ..., T$

$$\frac{n\eta}{4}\sum_{t=1}^{T}\|\nabla f(x_0^t)\|^2 \leq f(x_0^1) + L^2\eta^3 n^3 \sigma^2 T. \tag{74}$$

Divide $\frac{n\eta T}{4}$ on both sides

$$\min_{t=1,\ldots T} \|\nabla f(x_0^t)\|^2 \leq \frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_0^t)\|^2 \leq \frac{4(f(x_0) - f(x^*)))}{n\eta T} + 4L^2\eta^2 n^2 \sigma^2. \tag{75}$$

When $\sigma = 0$, substitute $\eta = \frac{1}{2nL\sqrt{\rho}}$ to obtain the result for SGC. ■

D.4. Sample Complexity of Strongly-Convex Objective

Corollary 3 *Suppose Assumptions 1,3 hold. Assume f is μ-strongly convex. Set $\eta = min\{\frac{1}{2nL\sqrt{\rho}}, \frac{2}{n\mu T}\log(\frac{(f(x_0)-f(x^*))\mu^3T^2}{L^2\sigma^2})\}$. Then*

$$f(x_n^T) - f(x^*) \leq \tilde{\mathcal{O}}\Big(\exp\big(\frac{-T}{\kappa\sqrt{\rho}}\big)\big(f(x_0) - f(x^*)\big) + \frac{\kappa^2\sigma^2}{\mu T^2}\Big). \tag{76}$$

Proof First consider $\eta = \frac{1}{2nL\sqrt{\rho}} \leq \frac{2}{n\mu T}\log(\frac{(f(x_0)-f(x^*))\mu^3T^2}{L^2\sigma^2})$

$$\begin{aligned} f(x_n^T) - f(x^*) &\leq (1 - \frac{1}{4\kappa\sqrt{\rho}})^T(f(x_0) - f(x^*)) + \frac{\sigma^2}{4\mu\rho} \\ &\leq \exp(\frac{-T}{4\kappa\sqrt{\rho}})(f(x_0) - f(x^*)) + \frac{8L^2\sigma^2}{\mu^3T^2}\log^2(\frac{f(x_0) - f(x^*)\mu^3T^2}{L^2\sigma^2}) \\ &= \tilde{\mathcal{O}}\Big(\exp\big(\frac{-T}{\kappa\sqrt{\rho}}\big)\big(f(x_0) - f(x^*)\big) + \frac{\kappa^2\sigma^2}{\mu T^2}\Big). \end{aligned} \tag{77}$$

For $\eta = \frac{2}{n\mu T}\log(\frac{(f(x_0)-f(x^*))\mu^3T^2}{L^2\sigma^2}) \leq \frac{1}{2\sqrt{2}nL\sqrt{\rho}}$

$$\begin{aligned} f(x_n^T) - f(x^*) &\leq \exp(-\frac{n\mu T}{2}\frac{2}{n\mu T}\log(\frac{(f(x_0) - f(x^*))\mu^3T^2}{L^2\sigma^2}))(f(x_0) - f(x^*)) + \\ &\quad \frac{2L^2n^2\sigma^2}{\mu}\frac{4}{n^2\mu^2T^2}\log^2(\frac{(f(x_0) - f(x^*))\mu^3T^2}{L^2\sigma^2}) \\ &= \frac{L^2\sigma^2}{\mu^3T^2} + \frac{8L^2\sigma^2}{\mu^3T^2}\log^2(\frac{(f(x_0) - f(x^*))\mu^3T^2}{L^2\sigma^2}) \\ &= \tilde{\mathcal{O}}(\frac{\kappa^2\sigma^2}{\mu T^2}). \end{aligned} \tag{78}$$

Combine (77) and (78) together

$$f(x_n^T) - f(x^*) \leq \tilde{\mathcal{O}}\Big(\exp\big(\frac{-T}{\kappa\sqrt{\rho}}\big)\Delta_0 + \frac{\kappa^2\sigma^2}{\mu T^2}\Big), \tag{79}$$

where $\Delta_0 = f(x_0) - f(x^*)$. Apply Lemma 2 to (79)

$$\|x_n^T - x^*\|^2 = \tilde{\mathcal{O}}\Big(\exp\big(-\frac{T}{\kappa\sqrt{\rho}}\big)\frac{\Delta_0}{\mu} + \frac{\kappa^2\sigma^2}{\mu^2T^2}\Big). \tag{80}$$

This translates to a sample complexity of $\tilde{\mathcal{O}}(n\kappa\sqrt{\rho} + \frac{n\kappa\sigma}{\mu\sqrt{\epsilon}})$ where logarithmic factors are ignored. ■

D.5. Sample Complexity of Non-Convex Objective

Corollary 4 *Suppose Assumptions 1,3 hold. Assume f is Non-Convex. Then the sample complexity of reach an ϵ-accurate solution ($\|\nabla f(x)\| \leq \epsilon$) is*

$$\tilde{\mathcal{O}}(\frac{L\sqrt{n}\sqrt{\rho+1+n}}{\epsilon^2} + \frac{L\sqrt{n}\sigma}{\epsilon^3}). \tag{81}$$

Proof Start from (75) in the Proof of Theorem 6

$$\min_{t=1,\ldots T} \|\nabla f(x_0^t)\|^2 \leq \frac{4(f(x_0) - f(x^*)))}{n\eta T} + 4L^2\eta^2 n^2 \sigma^2. \tag{82}$$

We first require $4L^2\eta^2 n^2\sigma^2 \leq \frac{\epsilon^2}{2}$, this implies

$$\eta \leq \frac{\epsilon}{2\sqrt{2}Ln\sigma}. \tag{83}$$

Note that for (82) to hold. We require $\eta \leq \frac{1}{2\sqrt{2}nL\sqrt{\rho}}$, combine this with (83), we have

$$\eta \leq \min\{\frac{1}{2\sqrt{2}nL\sqrt{\rho}}, \frac{\epsilon}{2\sqrt{2}Ln\sigma}\}. \tag{84}$$

Note that $\frac{4\mathbb{E}[f(x_0)-f(x^*)]}{n\eta T} \leq \frac{\epsilon^2}{2} \Rightarrow \frac{8[f(x_0^t - f(x^*))]}{\eta\epsilon^2} \leq nT$, this implies

$$\begin{aligned} nT \geq \frac{8[f(x_0^t - f(x^*))]}{\eta\epsilon^2} &= \frac{8[f(x_0^t) - f(x^*))]}{\epsilon^2} \max\{2\sqrt{2}nL\sqrt{\rho}, \frac{2\sqrt{2}Ln\sigma}{\epsilon}\} \\ &= \frac{16\sqrt{2}[f(x_0^t) - f(x^*)]Ln}{\epsilon^2} \max\{\sqrt{\rho}, \frac{\sigma}{\epsilon}\} \\ &= \tilde{\mathcal{O}}(\frac{Ln\sqrt{\rho}}{\epsilon^2} + \frac{Ln\sigma}{\epsilon^3}). \end{aligned} \tag{85}$$

■