Fast Convergence Rates of Random Reshuffling under Interpolation and the Polyak-Łojasiewicz Condition

Chen Fan FANCHEN2@MAIL.CS.UBC.CA
Christos Thrampoulidis CTHRAMPO@ECE.UBC.CA
Mark Schmidt SCHMIDTM@CS.UBC.CA
University of British Columbia, Canada

Abstract

Modern machine learning models are often over-parameterized and as a result they can interpolate the training data. Under such a scenario, we study the convergence properties of a sampling-without-replacement variant of Stochastic Gradient Descent (SGD), known as Random Reshuffling (RR). Unlike SGD that samples data with replacement at every iteration, RR chooses a random permutation of data at the beginning of each epoch. For under-parameterized models, it has been recently shown that RR converges faster than SGD when the number of epochs is larger than the condition number (κ) of the problem under standard assumptions like strong convexity. However, previous works do not show that RR outperforms SGD under interpolation for strongly convex objectives. Here, we show that for the class of Polyak-Łojasiewicz (PL) functions that generalizes strong convexity, RR can outperform SGD as long as the number of samples (n) is less than the parameter (ρ) of a strong growth condition (SGC).

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. 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, $\mathcal{O}(\frac{1}{k})$ 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 a linear convergence rate like full gradient descent under various interpolation conditions for strongly-convex functions (Moulines and Bach, 2011, Needell et al., 2014, Schmidt and Roux, 2013). An assumption that is weaker than strong convexity which allows full gradient descent to achieve a linear rate is the Polyak-Łojasiewicz (PL) condition (Polyak, 1963). Recently, the PL condition has gained

© C. Fan, C. Thrampoulidis & M. Schmidt.

popularity in machine learning (Karimi et al., 2016) and it has been shown that several overparameterized models that interpolate the data satisfy the PL condition (Bassily et al., 2018, Oymak and Soltanolkotabi, 2019, Soltanolkotabi et al., 2018). Notably, under interpolation and the PL condition SGD can achieve a linear rate similar to full gradient descent (Bassily et al., 2018, Vaswani et al., 2019). This paper shows that further convergence gains can be achieved under the PL condition and an interpolation assumption when a sampling-without-replacement counterpart of SGD, known as Random Reshuffling (RR) is used.

RR has long been known to converge faster than SGD empirically for certain problems (Bottou, 2009, 2012). However, analyzing RR is more difficult than SGD because (conditioned on the past iterates) each individual gradient is no longer an unbiased estimate of the full gradient. Thus, the analysis of RR has only emerged in a series of recent efforts (Gürbüzbalaban et al., 2021, Haochen and Sra, 2019, Nagaraj et al., 2019, Safran and Shamir, 2020). Previous works have shown that RR outperforms SGD for strongly-convex objectives in various under-parameterized settings, when the the number of epochs (T) is larger than the condition number of the problem (κ). However, in the over-parameterized settings current convergence rate analyses do not show that RR is faster than SGD (see Table 1). In this work, we address these questions by analyzing RR (and IG) for PL functions under the strong growth condition (SGC) interpolation condition (Schmidt and Roux, 2013). Under the SGC, we give an explicit convergence rate for RR (and IG) that can be faster than the best known rate for SGD. These advantages of RR do not exist for current analyses of convex or strongly-convex functions.

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 the 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. $\pi_{j+1}^t = j + 1$.

Table 1: Number of gradient evaluations of each algorithm to obtain an ϵ-accurate solution, which is defined as $\|x - x^*\|^2 \leq \epsilon$ and $f(x) - f(x^*) \leq \epsilon$ for μ-strongly convex and μ-PL objectives respectively. [1] All results hold under interpolation.

Citation	Algorithm	μ-Strongly Convex	μ-PL
Needell et al. (2014)	SGD	$\frac{L_{\max}}{\mu}$	-
Vaswani et al. (2019)[2]	SGD	$\alpha\frac{L}{\mu}$	-
Mishchenko et al. (2020)[3]	IG	$n\frac{L_{\max}}{\mu}$	-
	RR	$n\frac{L_{\max}}{\mu}$	-
Bassily et al. (2018)	SGD	-	$\frac{L_{\max}^2}{\mu^2}$
Vaswani et al. (2019)	SGD	-	$\rho\frac{L}{\mu}$
Nguyen et al. (2021)[4]	IG	-	$\frac{n}{\sqrt{\epsilon}}$
	RR	-	$\frac{\sqrt{n}}{\sqrt{\epsilon}}$
This work (SGC)	IG	-	$n\sqrt{\rho}\frac{L_{\max}}{\mu}$
	RR	-	$\sqrt{n}\sqrt{\rho + 1 + n}\frac{L_{\max}}{\mu}$

[1] We ignore numerical constants and logarithmic factors.

[2] For μ-strongly convex objectives, Vaswani et al. (2019) assume a weak growth condition (WGC), i.e. $\frac{1}{n}\sum_{i=1}^n \|\nabla f(x; i)\| \leq 2\rho L[f(x) - f(x^*)]$; for μ-PL, they assume the SGC. Note that $L \leq L_{\max}$, $\alpha \leq \frac{L_{\max}}{L}$, and $\rho \leq \frac{L_{\max}}{\mu}$ (Mishkin, 2020).

[3] Mishchenko et al. (2020) assume that each $f(\cdot; i)$ is convex. We do not make this assumption.

[4] The dependence on ϵ outside of a logarithmic factor in the results of Nguyen et al. (2021) yields slower sublinear convergence rates.

1.2. Assumptions

In this section, we present the assumptions made in our analyses. First, we give the definition of the Polyak-Łojasiewicz (PL) inequality

Assumption 1 (μ-PL) *There exists some $\mu > 0$ such that*

$$f(x) - f^* \leq \frac{1}{2\mu}\|\nabla f(x)\|^2 \quad \forall x \in \mathrm{dom}(f), \tag{3}$$

where f^* is the optimal value of f.

PL implies that every stationary point of f is a global minimum, but is weaker than strong convexity. The textbook example of a non-strongly convex function that satisfies PL is that of least-squares (see Karimi et al., 2016). More recent literature shows that PL condition holds for more complex over-parameterized models, including several classes of neural networks that interpolate under square loss (Soltanolkotabi et al., 2018). Second, we assume that the individual functions are smooth.

Assumption 2 *The objective f is L-smooth and each individual loss $f(\cdot; i)$ is L_i-smooth such that $\forall x, x' \in \mathrm{dom}(f)$*

$$\|\nabla f(x; i) - \nabla f(x'; i)\| \leq L_i\|x - x'\|. \tag{4}$$

Denote $L_{\max} = \max_i L_i$. We assume f is lower bounded by f^*, which is achieved at some x^*, so $f^* = f(x^*)$. We also assume that each $f(\cdot; i)$ is lower bounded by some f_i^*.

Next, we formally define interpolation.

Assumption 3 *We are in the interpolating regime, which we take to mean that*

$$\nabla f(x^*) = 0 \quad \Longrightarrow \quad \nabla f(x^*; i) = 0. \tag{5}$$

Thus, by interpolation we mean that stationary points with respect to the function f are also stationary points with respect to the individual functions $f(\cdot; i)$. Finally, we introduce the strong growth condition, which is stronger than interpolation as it implies Assumption 3.

Assumption 4 (SGC) *There exists a constant $\rho \geq 1$ such that the following holds*

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x; i)\|^2 \leq \rho\|\nabla f(x)\|^2, \quad \forall x \in \mathrm{dom}(f). \tag{6}$$

For smooth functions in interpolating settings, the SGC is related to PL. Concretely: under interpolation, a smooth and PL objective f also satisfies the SGC with a ρ that is at most $\frac{L_{\max}}{\mu}$ (Vaswani et al., 2019, Prop. 2). There are also other interesting function classes, not necessarily PL, that satisfy the SGC. The proposition below extends (Vaswani et al., 2019, Lem. 1) to a function class that includes squared-hinge or logistic losses as special cases.

Proposition 1 *Assume binary linearly-separable dataset $(a_i, y_i), i \in [n]$ of size n with margin $\tau := \arg\max_{\|x\|_2=1} \min_{i\in[n]} y_i a_i^T x > 0$, normalized features $\|a_i\|_2 \leq 1$ and $y_i \in \{\pm 1\}$. Let $f(x; i) = \ell(y_i a_i^T x)$ for a smooth monotonic function ℓ. Then, SGC (6) holds with $\rho = n/\tau^2$.*

We also consider a relaxation of the SGC that does not require the data to be fit exactly (Cevher and Vũ, 2019, Polyak and Tsypkin, 1973).

Assumption 5 *There exists constants $\rho \geq 0$ and $\sigma \geq 0$ such that the following holds: $\forall x \in \mathrm{dom}(f)$,*

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x;i)\|^2 \leq \rho\|\nabla f(x)\|^2 + \sigma^2. \tag{7}$$

Assumption 5 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).

2. Related Work

Optimization under the PL condition The PL inequality (Assumption 1) was first explored by Polyak (1963) and Lojasiewicz (1963). It applies to a wide range of important machine learning problems such as least square and logistic regression (over a compact set) (Karimi et al., 2016). More generally, any function of the form $f(x) = g(Ax)$ for a matrix A with a g being μ-strongly convex satisfies the μ-PL condition. Some over-parametrized models such as deep neural networks may contain stationary points that are sub-optimal, which is incompatible with the PL assumption. Nevertheless, several works have argued that considering a local PL condition around minimizers can be used as a model for analyzing the effectiveness of SGD in training deep neural networks (Liu et al., 2022, Oymak and Soltanolkotabi, 2019). Polyak (1963) showed that full gradient descent can achieve linear convergence rate under the PL condition. But it has recently been highlighted that the PL condition can be used to show linear convergence rates of a variety of methods (Karimi et al., 2016). Typically, the PL condition leads to similar convergence rates as those obtained under the stronger condition of strong convexity. In the case of SGD under interpolation, it has been shown that the rate of SGD under the PL condition is linear (Bassily et al., 2018, Vaswani et al., 2019). However, the convergence rates for SGD under interpolation for μ-PL functions are slower than those for strongly convex functions (see Table 1).

RR for Under-Parameterized Problems The difficulty in the analysis of RR arises because of the bias in the conditional expectation of gradients, i.e. $\mathbb{E}\big[\nabla f(x_i^t;\pi_{i+1}^t) \mid x_i^t\big] \neq \nabla f(x_i^t)$. An early attempt to analyze RR by Recht and Ré (2012) was not successful because their noncommutative arithmetic-geometric mean inequality conjecture was proven to be false (Lai and Lim, 2020). The non-asymptotic convergence properties of RR have only been addressed recently. Haochen and Sra (2019) give the first convergence result of $\tilde{\mathcal{O}}(\frac{1}{n^2T^2} + \frac{1}{T^3})$, where T is the number of epochs. The rate of RR was improved by Nagaraj et al. (2019) to $\tilde{\mathcal{O}}(\frac{1}{nT^2})$ when $T \gtrsim \kappa^2$ by assuming 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 all i), which may require a constraint on problem (1) and a projection operation is needed to ensure the 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). Our analysis does not make the component-wise convexity or G-Lipschitzness assumption.

RR for Over-Parameterized Problems Despite the widespread use of RR for training over-parame terized models, there is relatively little literature analyzing this setting. Haochen and Sra (2019) show that the convergence rate of RR is at least as fast as SGD under interpolation even without any epoch requirements. However, the result of Haochen and Sra (2019) does not show that a gap in the rates can exist and only applies in the degenerate case where each function is strongly-convex.[1] We can also obtain results under interpolation as special cases of the results of Mishchenko et al. (2020) and Nguyen et al. (2021). However, in the interpolation setting the rates obtained by these works for convex and strongly-convex functions are slower than the rate obtained by Vaswani et al. (2019) for SGD (see Table 1). To our knowledge this is the first work to show RR can outperform SGD in an over-parameterized setting.

3. Contributions

Our main contributions are summarized as follows:

- For μ-PL functions satisfying the SGC, we derive the sample complexity of RR to be $\tilde{\mathcal{O}}(\frac{L_{\max}}{\mu}\sqrt{n}\sqrt{\rho+1+n})$. In comparison, the sample complexity of SGD in this case is $\tilde{\mathcal{O}}(\frac{L}{\mu}\rho)$. Hence, RR outperforms SGD when $\boldsymbol{\rho} \gg \boldsymbol{n}$ and $L_{\max} \sim L$ without requiring a large epoch. The situation $\boldsymbol{\rho} \gg \boldsymbol{n}$ happens when there is a large amount of disagreements in the gradients. This is in clear contrast with the strongly convex case where existing bounds for RR do *not* show any gains over SGD.

- While the choice of RR over SGD under the SGC crucially depends on the relative magnitude of ρ and n, we show that RR consistently outperforms IG for both small and large ρ. Moreover, IG can outperform SGD for the μ-PL case when $\boldsymbol{\rho} \gg \boldsymbol{n^2}$ and $L_{\max} \sim L$.

- We conduct experiments on the objectives considered by our theory. We verify that the condition $\boldsymbol{\rho} \gg \boldsymbol{n}$ 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.

4. Theory Results

We present the convergence results of RR and IG for μ-PL objectives under SGC. Below, we use a constant step size η. When comparing our results to SGD, we assume $L_{\max} \sim L$, and denote $\kappa = \frac{L_{\max}}{\mu} \sim \frac{L}{\mu}$.

1. This setting is uninteresting under interpolation because we could solve the problem by simply applying gradient descent to any individual function and ignoring all other training examples.

Theorem 1 (μ-PL + RR) *Suppose Assumptions 1,2,5 hold, and choose a step size* $\eta \leq \frac{1}{2\sqrt{2}L_{\max}\sqrt{n(\rho+1+n)}}$. *Then, we have for* RR *that*

$$\mathbb{E}[f(x_n^T) - f(x^*)] \leq \big(1 - \frac{1}{2}n\mu\eta\big)^T\big(f(x_0) - f^*\big) + \frac{4L_{\max}^2\eta^2 n\sigma^2}{\mu}. \tag{8}$$

Further assuming SGC (Assumption 4), we obtain for $\eta = \frac{1}{2\sqrt{2}L_{\max}\sqrt{n(\rho+1+n)}}$ *that*

$$\mathbb{E}[f(x_n^T) - f(x^*)] \leq \big(1 - \frac{\sqrt{n}}{4\sqrt{2}\kappa\sqrt{\rho+1+n}}\big)^T\big(f(x_0) - f^*\big). \tag{9}$$

With a proper choice of η, we can in fact translate (8) into a sample complexity of $\tilde{\mathcal{O}}(\kappa\sqrt{n}\sqrt{\rho+1+n} + \frac{\kappa\sqrt{n}\sigma}{\sqrt{\mu\epsilon}})$; see Corollary 1 in Appendix. Comparing this with $\tilde{\mathcal{O}}(\kappa\rho + \frac{\kappa\sqrt{\rho}\sigma}{\sqrt{\mu\epsilon}})$ 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.

Theorem 2 (μ-PL + IG) *Suppose Assumptions 1,2,5 hold, and choose a step size* $\eta \leq \frac{1}{2nL_{\max}\sqrt{\rho}}$. *Then, we have for IG that*

$$f(x_n^T) - f^* \leq (1 - \frac{1}{2}n\mu\eta)^T(f(x_0) - f^*) + \frac{2L_{\max}^2\eta^2 n^2\sigma^2}{\mu}. \tag{10}$$

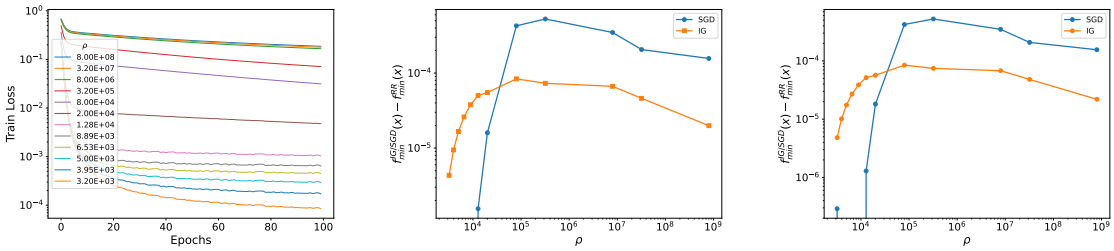
Further assuming SGC (Assumption 4), set $\eta = \frac{1}{2nL_{\max}\sqrt{\rho}}$ *to obtain*

$$f(x_n^T) - f^* \leq (1 - \frac{1}{4\kappa\sqrt{\rho}})^T(f(x_0) - f^*). \tag{11}$$

The sample complexity of IG for the μ-PL case is $\tilde{\mathcal{O}}(\kappa n\sqrt{\rho} + \frac{\kappa n\sigma}{\sqrt{\mu\epsilon}})$; see Corollary 2 in the Appendix. Under SGC, this is worse than RR whether $n < \rho$ or $n \geq \rho$. IG is also worse than RR in the case of $\sigma \neq 0$ due to the extra factor of $\sqrt{n}$ in the 2nd term of the sample complexity. On the other hand, IG can outperform SGD under SGC, provided that $n^2 \ll \rho$. *In summary, under SGC and PL,* RR *always outperforms IG, and also, it outperforms SGD in ill-conditioned or low-sample regimes* $n \ll \rho$. *Moreover, when* $n^2 \ll \rho$ *IG outperforms SGD.*

5. Experimental Results

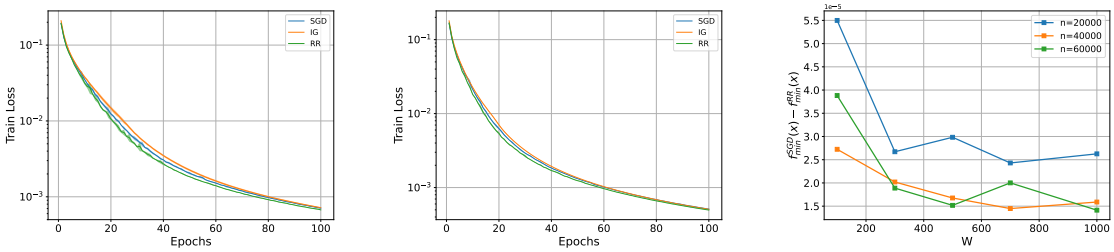
We first conduct experiments on a binary classification synthetic dataset. Specifically, we use the code provided by Loizou et al. (2021) to generate linearly-separable data with margin $\tau > 0$. For training, we use a squared hinge loss with or without L_2 regularization to simulate the convex and μ-PL cases respectively. Note that squared hinge loss satisfies the SGC with $\rho = \frac{n}{\tau^2}$ based on Proposition 1. 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).



(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.

Besides understanding the influence of ρ, we further conduct experiments to study the convergence rates of RR in non-convex settings under interpolation. To this end, we use a Softmax loss and one hidden-layer multilayer perceptrons (1-MLP) of different widths on the MNIST image classification dataset (LeCun et al., 1998). 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 ($W = 100$) (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 $n = 20000$.

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.

Raef Bassily, Mikhail Belkin, and Siyuan Ma. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

Léon Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the symposium on learning and data science, Paris*, volume 8, pages 2624–2633, 2009.

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.

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

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.

Zehua Lai and Lek-Heng Lim. Recht-ré noncommutative arithmetic-geometric mean conjecture is false. In *International Conference on Machine Learning*, pages 5608–5617. PMLR, 2020.

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.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

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.

Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

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

Aaron Mishkin. *Interpolation, growth conditions, and stochastic gradient descent*. PhD thesis, University of British Columbia, 2020.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

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.

Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Advances in neural information processing systems*, 27, 2014.

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.

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.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960. PMLR, 2019.

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

BT Polyak and Ya Z Tsypkin. Pseudogradient adaptation and training algorithms. *Automation and remote control*, 34:45–67, 1973.

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.

Benjamin Recht and Christopher Ré. Toward a noncommutative arithmetic-geometric mean inequality: Conjectures, case-studies, and consequences. In *Conference on Learning Theory*, pages 11–1. JMLR Workshop and Conference Proceedings, 2012.

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.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

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.

Appendix A. Key Lemmas

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 2), 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 3 and Lemma 4). In this section, we present these lemmas that will be used in the theory proofs. Lemma 1 is a restatement of (Mishchenko et al., 2020, Lemma 1).

Lemma 1 *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*

$$\mathbb{E}[\bar{X}_\pi] = \bar{X} \qquad \mathbb{E}[\|\bar{X}_\pi - \bar{X}\|^2] = \frac{n-k}{k(n-1)}\sigma^2. \tag{12}$$

Lemma 2 is proved in (Nguyen et al., 2021, Lemma 8). Here we replace L with $L_{\max}$ and take the step size η to be constant for all iterations and epochs.

Lemma 2 *Suppose Assumption 2 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:*

$$f(x_0^{t+1}) \leq f(x_0^t) - \frac{n\eta}{2}\|\nabla f(x_0^t)\|^2 + \frac{L_{\max}^2\eta}{2}\sum_{i=0}^n \|x_i^t - x_0^t\|^2. \tag{13}$$

Taking total expectation over the randomness

$$\mathbb{E}[f(x_0^{t+1})] \leq \mathbb{E}[f(x_0^t)] - \frac{n\eta}{2}\mathbb{E}[\|\nabla f(x_0^t)\|^2] + \frac{L_{\max}^2\eta}{2}\sum_{i=0}^n \mathbb{E}[\|x_i^t - x_0^t\|^2]. \tag{14}$$

The remaining Lemmas 3-4 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. Lemma 3 will be combined with Lemma 2 in the proofs of IG.

Lemma 3 *Suppose Assumptions 2, 5 hold. For a step size $\eta \leq \frac{1}{\sqrt{2}nL_{\max}}$, it holds for IG that*

$$\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2 \leq 2\eta^2 n^3 \rho \|\nabla f(x_0^t)\|^2 + 2\eta^2 n^3 \sigma^2. \tag{15}$$

Proof By the update rule

$$
\begin{aligned}
\|x_i^t - x_0^t\|^2 &= \eta^2 \|\sum_{j=0}^{i-1} \nabla f(x_j^t; j+1)\|^2 \\
&= \eta^2 \|\sum_{j=0}^{i-1} \nabla f(x_j^t; j+1) - \nabla f(x_0^t; j+1) + \nabla f(x_0^t; j+1)\|^2 \\
&\leq 2\eta^2 \|\sum_{j=0}^{i-1} \nabla f(x_j^t; j+1) - \nabla f(x_0^t; j+1)\|^2 + 2\eta^2 \|\sum_{j=0}^{i-1} \nabla f(x_0^t; j+1)\|^2 \\
&\leq 2\eta^2 i \sum_{j=0}^{i-1} \|\nabla f(x_j^t; j+1) - \nabla f(x_0^t; j+1)\|^2 + 2\eta^2 i \sum_{j=0}^{i-1} \|\nabla f(x_0^t; j+1)\|^2 \\
&\leq 2\eta^2 L_{\max}^2 i \sum_{j=0}^{i-1} \|x_j^t - x_0^t\|^2 + 2\eta^2 i \sum_{j=0}^{i-1} \|\nabla f(x_0^t; j+1)\|^2. \qquad (16)
\end{aligned}
$$

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

$$
\begin{aligned}
\sum_{i=0}^{n-1} \|x_i^t - x_0^t\|^2 &\leq 2\eta^2 L_{\max}^2 \sum_{j=0}^{i-1} \|x_j^t - x_0^t\|^2 \sum_{i=0}^{n-1} i + 2\eta^2 \sum_{j=0}^{n-1} \|\nabla f(x_0^t; j+1)\|^2 \sum_{i=0}^{n-1} i \\
&\leq \eta^2 L_{\max}^2 n^2 \sum_{j=0}^{n-1} \|x_j^t - x_0^t\|^2 + \eta^2 n^3 \frac{1}{n} \sum_{j=0}^{n-1} \|\nabla f(x_0^t; j+1)\|^2 \\
&\leq \eta^2 L_{\max}^2 n^2 \sum_{j=0}^{n-1} \|x_j^t - x_0^t\|^2 + \eta^2 n^3 (\rho \|\nabla f(x_0^t)\|^2 + \sigma^2) \\
&= \eta^2 L_{\max}^2 n^2 \sum_{j=0}^{n-1} \|x_j^t - x_0^t\|^2 + \eta^2 n^3 \rho \|\nabla f(x_0^t)\|^2 + \eta^2 n^3 \sigma^2. \qquad (17)
\end{aligned}
$$

Finally, choosing $\eta \leq \frac{1}{\sqrt{2} n L_{\max}}$ leads to

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

■

Lemma 4 below will be combined with Lemma 2 in the proofs of RR.

Lemma 4 *Suppose Assumption 2, 5 hold. For a step size $\eta \leq \frac{1}{\sqrt{3} n L_{\max}}$, the following holds for RR*

$$
\mathbb{E}\Big[\sum_{i=0}^{n-1} \|x_i^t - x_0^t\|^2\Big] \leq 4n^2\eta^2(\rho + 1 + n)\mathbb{E}[\|\nabla f(x_0^t)\|^2] + 4\eta^2 n^2 \sigma^2. \qquad (19)
$$

Proof By the update rule

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

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 1 to find that

$$
\begin{aligned}
\mathbb{E}\Big[\|\frac{1}{i} \sum_{j=0}^{i-1} \nabla f(x_0^t; \pi_{j+1}^t) - \nabla f(x_0^t)\|^2 | \sigma^t\Big] &= \frac{n-i}{i(n-1)} \frac{1}{n} \sum_{j=0}^{n-1} \|\nabla f(x_0^t; j+1) - \nabla f(x_0^t)\|^2 \\
&\leq \frac{n-i}{i(n-1)} \frac{1}{n} \sum_{j=0}^{n-1} [2\|\nabla f(x_0^t; j+1)\|^2 + 2\|\nabla f(x_0^t)\|^2].
\end{aligned} \tag{21}
$$

Take conditional expectation of (20) and substitute (21) back

$$
\begin{aligned}
\mathbb{E}\Big[\|x_i^t - x_0^t\| | \sigma^t\Big] &\leq 3\eta^2 i L_{\max}^2 \sum_{j=0}^{n-1} \mathbb{E}[\|x_j^t - x_0^t\| | \sigma^t] + 6\eta^2 \frac{i(n-i)}{(n-1)} \frac{1}{n} \sum_{j=0}^{n-1} \|\nabla f(x_0^t; j+1)\|^2 + \\
&\quad 6\eta^2 \frac{i(n-i)}{n-1} \|\nabla f(x_0^t)\|^2 + 3\eta^2 i^2 \|\nabla f(x_0^t)\|^2 \qquad (22) \\
&\leq 3\eta^2 i L_{\max}^2 \sum_{j=0}^{n-1} \mathbb{E}[\|x_j^t - x_0^t\| | \sigma^t] + 6\eta^2 \frac{i(n-i)}{n-1} (\rho \|\nabla f(x_0^t)\|^2 + \sigma^2) + \\
&\quad 6\eta^2 \frac{i(n-i)}{n-1} \|\nabla f(x_0^t)\|^2 + 3\eta^2 i^2 \|\nabla f(x_0^t)\|^2. \qquad (23)
\end{aligned}
$$

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

$$\mathbb{E}\Big[\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2\Big] \leq 3\eta^2 L_{\max}^2 \mathbb{E}[\sum_{j=0}^{n-1}\|x_j^t - x_0^t\|^2]\sum_{i=0}^{n-1} i + 6\eta^2\frac{1}{n-1}(\rho\mathbb{E}[\|\nabla f(x_0^t)\|^2] + \sigma^2)\sum_{i=0}^{n-1} i(n-i) +$$

$$6\eta^2\frac{1}{n-1}\mathbb{E}[\|\nabla f(x_0^t)\|^2]\sum_{i=0}^{n-1} i(n-i) + 3\eta^2\mathbb{E}[\|\nabla f(x_0^t)\|^2]\sum_{i=0}^{n-1} i^2$$

$$\leq \frac{3}{2}\eta^2 L_{\max}^2 n^2 \mathbb{E}[\sum_{j=0}^{n-1}\|x_j^t - x_0^t\|^2] + 2\eta^2 n^2 \rho\mathbb{E}[\|\nabla f(x_0^t)\|^2] + 2\eta^2 n^2\sigma^2 +$$

$$2\eta^2 n^2\mathbb{E}[\|\nabla f(x_0^t)\|^2] + 2\eta^2 n^3\mathbb{E}[\|\nabla f(x_0^t)\|^2] \tag{24}$$

$$= \frac{3}{2}\eta^2 L_{\max}^2 n^2 \mathbb{E}[\sum_{j=0}^{n-1}\|x_j^t - x_0^t\|^2] + 2\eta^2 n^2(\rho + 1 + n)\mathbb{E}[\|\nabla f(x_0^t)\|^2] + 2\eta^2 n^2\sigma^2. \tag{25}$$

In (24), 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 $\eta \leq \frac{1}{\sqrt{3}L_{\max}n}$, we have

$$\mathbb{E}[\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2] \leq 4\eta^2 n^2(\rho + 1 + n)\mathbb{E}[\|\nabla f(x_0^t)\|^2] + 4\eta^2 n^2\sigma^2. \tag{26}$$

■

Appendix B. Convergence Proofs

B.1. Proof of Proposition 1

Proof We consider smooth monotonic non-increasing functions of the form $f_i(x) = l(y_i a_i^T x) = l(z_i^T x)$, where $z_i = y_i a_i$ and a_i is the feature vector for the ith sample. We assume the values of z_i are properly normalized such that $\max_i \|z_i\| \leq 1$. Define $x^* = \operatorname{argmax}_{\|x\|_2=1} \min_i z_i^t x$ and $\tau = \max_{\|x\|_2=1} \min_i z_i^T x$. Then we have

$$\|\frac{1}{n}\sum_{i=1}^{n}\nabla f(x;i)\|_2^2 = \|\frac{1}{n}\sum_{i=1}^{n} l^{'}(z_i^T x) z_i\|_2^2. \tag{27}$$

(28)

Note that for a vector u, its 2-norm is $\|u\|_2 = \max_{\|v\|_2=1} v^T u$. Hence, we have the following

$$
\begin{aligned}
\|\frac{1}{n}\sum_{i=1}^{n}\nabla f(x;i)\|_2^2 &\geq \big(\frac{1}{n}\sum_{i=1}^{n} l^{'}(z_i^T x) z_i^T x^*\big)^2 \\
&= \tau^2\big(\frac{1}{n}\sum_{i=1}^{n} l^{'}(z_i^T x)\big)^2 \\
&= \tau^2\Big\{\frac{1}{n^2}\sum_{i=1}^{n} l^{'}(z_i^T x)^2 + \frac{1}{n^2}\sum_{i\neq j} l^{'}(z_i^T x) l^{'}(z_j^T x)\Big\} \\
&\geq \frac{\tau^2}{n}\Big(\frac{1}{n}\sum_{i=1}^{n} l^{'}(z_i^T x)^2\Big) \qquad (29) \\
&\geq \frac{\tau^2}{n}\Big(\frac{1}{n}\sum_{i=1}^{n} l^{'}(z_i^T x)^2 \|z_i\|^2\Big) \\
&= \frac{\tau^2}{n}\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x;i)\|^2, \qquad (30)
\end{aligned}
$$

where the inequality in (29) follows because l is monotonic, that is $l^{'}(t_1)l^{'}(t_2) \geq 0\ \forall t_1, t_2 \in \mathbb{R}$. Therefore, we have the following

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x;i)\|^2 &\leq \frac{n}{\tau^2}\|\frac{1}{n}\sum_{i=1}^{n}\nabla f(x;i)\|_2^2 \\
&= \rho\|\nabla f(x)\|^2, \qquad (31)
\end{aligned}
$$

where $\rho = \frac{n}{\tau^2}$. ■

B.2. Proof of Theorem 1

Apply Lemma 2 and Lemma 4

$$
\begin{aligned}
\mathbb{E}[f(x_0^{t+1})] &\leq \mathbb{E}[f(x_0^t)] - \frac{n\eta}{2}\mathbb{E}[\|\nabla f(x_0^t)\|^2] + \frac{L_{\max}^2\eta}{2}[4\eta^2 n^2(\rho+1+n)\mathbb{E}[\|\nabla f(x_0^t)\|^2] + 4\eta^2 n^2\sigma^2] \\
&= \mathbb{E}[f(x_0^t)] - \frac{n\eta}{2}[1 - 4L_{\max}^2\eta^2 n(\rho+1+n)]\mathbb{E}[\|\nabla f(x_0^t)\|^2] + 2L_{\max}^2\eta^3 n^2\sigma^2. \qquad (32)
\end{aligned}
$$

Further use the PL inequality in Assumption 1 to arrive at

$$
\begin{aligned}
\mathbb{E}[f(x_0^{t+1}) - f(x^*)] \leq \mathbb{E}[f(x_0^t) - f(x^*)] - n\eta\mu[1 - 4L_{\max}^2\eta^2 n(\rho+1+n)]\mathbb{E}[f(x_0^t) - f(x^*)] + \\
2L_{\max}^2\eta^3 n^2\sigma^2. \qquad (33)
\end{aligned}
$$

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

$$
\mathbb{E}[f(x_0^{t+1}) - f(x^*)] \leq (1 - \frac{1}{2}n\mu\eta)\mathbb{E}[f(x_0^t) - f(x^*)] + 2L_{\max}^2\eta^3 n^2\sigma^2. \qquad (34)
$$

Solving (34) recursively for $t = 1, ..., T$ gives

$$\mathbb{E}[f(x_0^{T+1}) - f(x^*)] \leq (1 - \frac{1}{2}n\mu\eta)^T (f(x_0^1) - f(x^*)) + 2L_{\max}^2 \eta^3 n^2 \sigma^2 \sum_{j=0}^{T} (1 - \frac{1}{2}n\mu\eta)^j. \quad (35)$$

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

$$\begin{aligned}
\mathbb{E}[f(x_n^T) - f(x^*)] &\leq (1 - \frac{1}{2}n\mu\eta)^T (f(x_0^1) - f(x^*)) + 2L_{\max}^2 \eta^3 n^2 \sigma^2 \sum_{j=0}^{\infty} (1 - \frac{1}{2}n\mu\eta)^j \\
&\leq (1 - \frac{1}{2}n\mu\eta)^T (f(x_0) - f(x^*)) + 2L_{\max}^2 \eta^3 n^2 \sigma^2 (\frac{2}{n\mu\eta}) \\
&= (1 - \frac{1}{2}n\mu\eta)^T (f(x_0) - f(x^*)) + \frac{4L_{\max}^2 \eta^2 n \sigma^2}{\mu}. \quad (36)
\end{aligned}$$

Note that for (36) to hold, we require $\eta \leq \min\{\frac{1}{2\sqrt{2}L_{\max}\sqrt{n(\rho+1+n)}}, \frac{1}{\sqrt{3}L_{\max}n}\}$ because of Lemma 4. This condition becomes $\eta \leq \frac{1}{2\sqrt{2}L_{\max}\sqrt{n(\rho+1+n)}}$ as $\frac{1}{2\sqrt{2}L_{\max}\sqrt{n(\rho+1+n)}} \leq \frac{1}{\sqrt{3}L_{\max}n}$. Under SGC, substitute $\sigma = 0$ with $\eta = \frac{1}{2\sqrt{2}L_{\max}\sqrt{n(\rho+1+n)}}$ to obtain the desired result.

B.3. Sample Complexity for Strongly-Convex Objective under RR

Corollary 1 *Suppose Assumptions 1,2,5 hold. Assume f is μ-strongly convex, choose $\eta = \min\{\frac{1}{2\sqrt{2}L_{\max}\sqrt{n(\rho+1+n)}}, \frac{2}{n\mu T}\log(\frac{(f(x_0)-f(x^*))\mu^3 T^2 n}{L_{\max}^2\sigma^2})\}$, and define $\kappa \triangleq \frac{L_{max}}{\mu}$. Then we have for* RR

$$\mathbb{E}[f(x_n^T) - f(x^*)] = \tilde{\mathcal{O}}\Big(\exp\big(-\frac{T\sqrt{n}}{\kappa\sqrt{\rho+1+n}}\big)\big(f(x_0) - f(x^*)\big) + \frac{\kappa^2\sigma^2}{n\mu T^2}\Big). \quad (37)$$

Proof For the case $\eta = \frac{1}{2\sqrt{2}L_{\max}\sqrt{n(\rho+1+n)}} \leq \frac{2}{n\mu T}\log(\frac{(f(x_0)-f(x^*))\mu^3 T^2 n}{L_{\max}^2\sigma^2})$, substitute this η into (36)

$$\begin{aligned}
\mathbb{E}[f(x_n^T) - f(x^*)] &\leq (1 - \frac{\sqrt{n}}{4\sqrt{2}\kappa\sqrt{\rho+1+n}})^T (f(x_0) - f(x^*)) + \frac{\sigma^2}{2\mu(\rho+1+n)} \\
&\leq \exp(-\frac{T\sqrt{n}}{4\sqrt{2}\kappa\sqrt{\rho+1+n}})(f(x_0) - f(x^*)) + \frac{16L_{\max}^2\sigma^2}{n\mu^3T^2}\log^2(\frac{(f(x_0) - f(x^*))\mu^3T^2n}{L_{\max}^2\sigma^2}) \\
&= \tilde{\mathcal{O}}\Big(\exp\big(-\frac{T\sqrt{n}}{\kappa\sqrt{\rho+1+n}}\big)\big(f(x_0) - f(x^*)\big) + \frac{\kappa^2\sigma^2}{n\mu T^2}\Big). \quad (38)
\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}[f(x_n^T)-f(x^*)] &\leq \exp(-\frac{n\mu T}{2}\frac{2}{n\mu T}\log(\frac{((f(x_0)-f(x^*)))\mu^3T^2n}{L_{\max}^2\sigma^2})(f(x_0)-f(x^*))+ \\ &\quad \frac{4L_{\max}^2 n\sigma^2}{\mu}\frac{4}{n^2\mu^2T^2}\log^2(\frac{(f(x_0)-f(x^*))\mu^3T^2n}{L_{\max}^2\sigma^2}) \\ &= \frac{L_{\max}^2\sigma^2}{\mu^3T^2n}+\frac{16L_{\max}^2\sigma^2}{\mu^3T^2n}\log^2(\frac{(f(x_0)-f(x^*))\mu^3T^2n}{L_{\max}^2\sigma^2}) \\ &= \tilde{\mathcal{O}}(\frac{\kappa^2\sigma^2}{n\mu T^2}). \end{aligned} \tag{39}$$

Combining (38) and (39), we obtain

$$\mathbb{E}[f(x_n^T)-f(x^*)] = \tilde{\mathcal{O}}\Big(\exp\big(-\frac{T\sqrt{n}}{\kappa\sqrt{\rho+1+n}}\big)\Delta_0+\frac{\kappa^2\sigma^2}{n\mu T^2}\Big), \tag{40}$$

where $\Delta_0 = f(x_0)-f(x^*)$. This translates to a sample complexity of $\tilde{\mathcal{O}}(\kappa\sqrt{n}\sqrt{\rho+1+n}+\frac{\kappa\sqrt{n}\sigma}{\sqrt{\mu\epsilon}})$ where logarithmic factors are ignored. ■

B.4. Proof of Theorem 2

Proof Apply Lemma 2 and Lemma 3

$$\begin{aligned} f(x_0^{t+1}) &\leq f(x_0^t)-\frac{n\eta}{2}\|\nabla f(x_0^t)\|^2+\frac{L_{\max}^2\eta}{2}(2\eta^2n^3\rho\|\nabla f(x_0^t)\|^2+2\eta^2n^3\sigma^2) \\ &= f(x_0^t)-\frac{n\eta}{2}(1-2L_{\max}^2\eta^2n^2\rho)\|\nabla f(x_0^t)\|^2+L_{\max}^2\eta^3n^3\sigma^2. \end{aligned} \tag{41}$$

Choose $\eta \leq \frac{1}{2nL_{\max}\sqrt{\rho}}$ and apply PL inequality in Assumption 1

$$f(x_0^{t+1})-f(x^*) \leq (1-\frac{1}{2}n\mu\eta)(f(x_0^t)-f(x^*))+L_{\max}^2\eta^3n^3\sigma^2. \tag{42}$$

Solve (42) recursively

$$f(x_0^{T+1})-f(x^*) \leq (1-\frac{1}{2}n\mu\eta)^T(f(x_0^1)-f(x^*))+L_{\max}^2\eta^3n^3\sigma^2\sum_{j=0}^{T}(1-\frac{1}{2}n\mu\eta)^j. \tag{43}$$

This implies

$$\begin{aligned} f(x_n^T)-f(x^*) &\leq (1-\frac{1}{2}n\mu\eta)^T(f(x_0)-f(x^*))+L_{\max}^2\eta^3n^3\sigma^2\sum_{j=0}^{\infty}(1-\frac{1}{2}n\mu\eta)^j \\ &\leq (1-\frac{1}{2}n\mu\eta)^T(f(x_0)-f(x^*))+L_{\max}^2\eta^3n^3\sigma^2(\frac{2}{n\mu\eta}) \\ &= (1-\frac{1}{2}n\mu\eta)^T(f(x_0)-f(x^*))+\frac{2L_{\max}^2\eta^2n^2\sigma^2}{\mu}. \end{aligned} \tag{44}$$

Note for (44) to hold, we require $\eta \leq \min\{\frac{1}{2nL_{\max}\sqrt{\rho}}, \frac{1}{\sqrt{2}nL_{\max}}\}$ as we used Lemma 3. This reduces to $\eta \leq \frac{1}{2nL_{\max}\sqrt{\rho}}$ provided $\rho \geq 1$. Under SGC, set $\sigma = 0$ and substitute $\eta = \frac{1}{2nL_{\max}\sqrt{\rho}}$, we obtain the desired result. ■

B.5. Sample Complexity of Strongly-Convex Objective under IG

Corollary 2 *Suppose Assumptions 1,2,5 hold. Assume f is μ-strongly convex, set $\eta = \min\{\frac{1}{2nL_{\max}\sqrt{\rho}}, \frac{2}{n\mu T}\log(\frac{(f(x_0)-f(x^*))\mu^3T^2}{L_{\max}^2\sigma^2})\}$, and define $\kappa \triangleq \frac{L_{\max}}{\mu}$. Then we have for IG*

$$f(x_n^T) - f(x^*) \leq \tilde{\mathcal{O}}\Big(\exp\big(\frac{-T}{\kappa\sqrt{\rho}}\big)\big(f(x_0) - f(x^*)\big) + \frac{\kappa^2\sigma^2}{\mu T^2}\Big). \tag{45}$$

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

$$\begin{aligned} f(x_n^T) - f(x^*) &\leq (1 - \frac{1}{4\kappa\sqrt{\rho}})^T(f(x_0) - f(x^*)) + \frac{\sigma^2}{4\mu\rho} \\ &\leq \exp(\frac{-T}{4\kappa\sqrt{\rho}})(f(x_0) - f(x^*)) + \frac{8L_{\max}^2\sigma^2}{\mu^3T^2}\log^2(\frac{f(x_0) - f(x^*)\mu^3T^2}{L^2\sigma^2}) \\ &= \tilde{\mathcal{O}}\Big(\exp\big(\frac{-T}{\kappa\sqrt{\rho}}\big)\big(f(x_0) - f(x^*)\big) + \frac{\kappa^2\sigma^2}{\mu T^2}\Big). \end{aligned} \tag{46}$$

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

$$\begin{aligned} f(x_n^T) - f(x^*) &\leq \exp(-\frac{n\mu T}{2}\frac{2}{n\mu T}\log(\frac{(f(x_0) - f(x^*))\mu^3T^2}{L_{\max}^2\sigma^2}))(f(x_0) - f(x^*)) + \\ &\quad \frac{2L_{\max}^2n^2\sigma^2}{\mu}\frac{4}{n^2\mu^2T^2}\log^2(\frac{(f(x_0) - f(x^*))\mu^3T^2}{L_{\max}^2\sigma^2}) \\ &= \frac{L_{\max}^2\sigma^2}{\mu^3T^2} + \frac{8L_{\max}^2\sigma^2}{\mu^3T^2}\log^2(\frac{(f(x_0) - f(x^*))\mu^3T^2}{L_{\max}^2\sigma^2}) \\ &= \tilde{\mathcal{O}}(\frac{\kappa^2\sigma^2}{\mu T^2}). \end{aligned} \tag{47}$$

Combine (46) and (47) together

$$f(x_n^T) - f(x^*) \leq \tilde{\mathcal{O}}\Big(\exp\big(\frac{-T}{\kappa\sqrt{\rho}}\big)\Delta_0 + \frac{\kappa^2\sigma^2}{\mu T^2}\Big), \tag{48}$$

where $\Delta_0 = f(x_0) - f(x^*)$. This translates to a sample complexity of $\tilde{\mathcal{O}}(n\kappa\sqrt{\rho} + \frac{n\kappa\sigma}{\sqrt{\mu\epsilon}})$ where logarithmic factors are ignored. ■