Demystifying SGD with Doubly Stochastic Gradients

Kyurae Kim [1] **Joohwan Ko** [2] **Yi-An Ma** [3] **Jacob R. Gardner** [1]

Abstract

Optimization objectives in the form of a sum of intractable expectations are rising in importance (*e.g.*, diffusion models, variational autoencoders, and many more), a setting also known as "finite sum with infinite data." For these problems, a popular strategy is to employ SGD with *doubly stochastic gradients* (doubly SGD): the expectations are estimated using the gradient estimator of each component, while the sum is estimated by subsampling over these estimators. Despite its popularity, little is known about the convergence properties of doubly SGD, except under strong assumptions such as bounded variance. In this work, we establish the convergence of doubly SGD with independent minibatching and random reshuffling under general conditions, which encompasses dependent component gradient estimators. In particular, for dependent estimators, our analysis allows fined-grained analysis of the effect correlations. As a result, under a per-iteration computational budget of $b \times m$, where b is the minibatch size and m is the number of Monte Carlo samples, our analysis suggests where one should invest most of the budget in general. Furthermore, we prove that random reshuffling (RR) improves the complexity dependence on the subsampling noise.

1. Introduction

Stochastic gradient descent (SGD; Robbins & Monro, 1951; Bottou, 1999; Nemirovski et al., 2009; Shalev-Shwartz et al., 2011) is the *de facto* standard for solving large scale optimization problems of the form of finite sums such as

$$\underset{\boldsymbol{x}\in\mathcal{X}\subseteq\mathbb{R}^d}{\text{minimize}} \left\{ F(\boldsymbol{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}) \right\}. \tag{1}$$

When n is large, SGD quickly converges to low-accuracy solutions by subsampling over components $f_1, \dots, f_n$. The properties of SGD on the finite sum class have received an immense amount of interest (Bottou et al., 2018) as it includes empirical risk minimization (ERM; Vapnik, 1991).

Unfortunately, for an emerging large set of problems in machine learning, we may not have direct access to the components $f_1, \dots, f_n$. That is, each f_i may be defined as an intractable expectation, or an "infinite sum"

$$f_i(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\eta}\sim\varphi} f_i(\boldsymbol{x}; \boldsymbol{\eta}), \tag{2}$$

where we only have access to the noise distribution φ and the integrand $f_i(\boldsymbol{x}; \boldsymbol{\eta})$, and $\boldsymbol{\eta}$ is a potentially continuous and unbounded source of stochasticity; a setting Zheng & Kwok (2018); Bietti & Mairal (2017) have previously called "finite sum with infinite data." Such problems include the training of diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song & Ermon, 2019), variational autoencoders (Kingma & Welling, 2014; Rezende et al., 2014), solving ERM under differential privacy (Bassily et al., 2014; Song et al., 2013), and also classical problems such as variational inference (Ranganath et al., 2014; Titsias & Lázaro-Gredilla, 2014; Kucukelbir et al., 2017), and variants of empirical risk minimization (Dai et al., 2014; Bietti & Mairal, 2017; Shi et al., 2021; Orvieto et al., 2023; Liu et al., 2021). In contrast to the finite sum setting where SGD has traditionally been applied, our problem takes the form of

$$\underset{\boldsymbol{x}\in\mathcal{X}\subseteq\mathbb{R}^d}{\text{minimize}} \left\{ F(\boldsymbol{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\eta}\sim\varphi} f_i(\boldsymbol{x}; \boldsymbol{\eta}) \right\}.$$

These optimization problems are typically solved using SGD with *doubly stochastic gradients* (doubly SGD; coined by Dai et al. 2014; Titsias & Lázaro-Gredilla 2014), so-called because, in addition to subsampling over f_i, stochastic estimates of each component f_i are used.

Previous studies have relied on strong assumptions to analyze doubly stochastic gradients. For instance, Kulunchakov & Mairal (2020); Bietti & Mairal (2017); Zheng & Kwok (2018) have **(i)** assumed that the variance of each component estimator is bounded by a constant, which contradicts componentwise strong convexity (Nguyen et al.,

[1]Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA, U.S.A. [2]KAIST, Daejeon, South Korea, Republic of [3]Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, CA, U.S.A.. Correspondence to: Kyurae Kim <kyrkim@seas.upenn.edu>.

Proceedings of the 41st *International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

2018) when $\mathcal{X} = \mathbb{R}^d$, **(ii)** or that the integrand $\nabla f_i(\boldsymbol{x}; \boldsymbol{\eta})$, is L-Lipschitz smooth "uniformly" over $\boldsymbol{\eta}$. That is, for any fixed $\boldsymbol{\eta}$ and i,

$$\|\nabla f_i(\boldsymbol{x}; \boldsymbol{\eta}) - \nabla f_i(\boldsymbol{y}; \boldsymbol{\eta})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

holds for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^2$. Unfortunately, this only holds for additive noise and is otherwise unrealizable when $\boldsymbol{\eta}$ has an unbounded support. Therefore, analyses relying on uniform smoothness obscure a lot of interesting behavior. Meanwhile, weaker assumptions such as expected smoothness (ES; Moulines & Bach, 2011; Gower et al., 2021b) have shown to be realizable even for complex gradient estimators (Domke, 2019; Kim et al., 2023). Therefore, a key question is how these ES-type assumptions propagate to doubly stochastic estimators. Among these, we focus on the expected residual (ER; Gower et al., 2019) condition.

Furthermore, in practice, certain applications of doubly SGD share the randomness $\boldsymbol{\eta}$ across the batch B. (See Section 2.2 for examples.) This introduces dependence between the gradient estimate for each component such that $\nabla f_i(\boldsymbol{x}; \boldsymbol{\eta}) \not\perp\!\!\!\perp \nabla f_j(\boldsymbol{x}; \boldsymbol{\eta})$ for $i, j \in B$. Little is known about the effect of this practice apart from some empirical results (Kingma et al., 2015). For instance, when m Monte Carlo samples of $\boldsymbol{\eta}$ and a minibatch of size b are used, what is the trade-off between m and b? To answer this question, we provide a theoretical analysis of doubly SGD that encompasses dependent gradient estimators.

Technical Contributions

- **Theorem 1**: For doubly stochastic estimators, we establish a general variance bound of the form of

$$\mathcal{O}\left(\frac{\frac{1}{n}\sum_{i=1}^n \sigma_i^2}{mb} + \rho\frac{\left(\frac{1}{n}\sum_{i=1}^n \sigma_i\right)^2}{m} + \frac{\tau^2}{b}\right),$$

  where σ_i^2 is the variance of the estimator of ∇f_i, $\rho \in [0, 1]$ is the correlation between the estimators, and τ^2 is the variance of subsampling.
- **Theorems 2 and 3**: Using the general variance bound, we show that a doubly stochastic estimator subsampling over correlated estimators satisfying the ER condition and the bounded variance (BV; Definition 2; bounded only on the solution set) condition equally satisfies the ER and BV conditions as well. This is sufficient to guarantee the convergence of doubly SGD on convex, quasar convex, and nonconvex smooth objectives.
- **Theorem 5**: Under similar assumptions, we also prove the convergence of doubly SGD with random reshuffling (doubly SGD-RR), instead of independent subsampling, on a strongly convex objective with strongly convex components.

Practical Insights

- **Should I invest in (increase) m or b?** When dependent gradient estimators are used, increasing m or b does not have the same impact on the gradient variance as the subsampling strategy also affects the resulting correlation between the estimators. Through Lemma 9, our analysis provides insight into this effect. In particular, we reveal that reducing subsampling variance also reduces Monte Carlo variances. Therefore, for a fixed budget $m \times b$, increasing b should always be preferred over increasing m.
- **Random Reshuffling Improves Complexity.** Our analysis of doubly SGD-RR reveals that, for strongly convex objectives, random reshuffling improves the iteration complexity of doubly SGD from $\mathcal{O}\left(\frac{1}{\epsilon}\sigma_{\text{mc}}^2 + \frac{1}{\epsilon}\sigma_{\text{sub}}^2\right)$ to $\mathcal{O}\left(\frac{1}{\epsilon}\sigma_{\text{mc}}^2 + \frac{1}{\sqrt{\epsilon}}\sigma_{\text{sub}}\right)$. Furthermore, for dependent gradient estimators, doubly SGD-RR is "super-efficient": for a batch taking $\Theta(mb)$ samples to compute, it achieves a n/b tighter asymptotic sample complexity compared to full-batch SGD.

2. Preliminaries

Notation We denote random variables (RVs) in serif (*e.g.*, x, $\boldsymbol{\mathsf{x}}$, $\boldsymbol{\mathsf{X}}$, B), vectors and matrices in bold (*e.g.*, $\boldsymbol{x}$, $\boldsymbol{\mathsf{x}}$, $\boldsymbol{A}$, $\boldsymbol{\mathsf{A}}$). For a vector $\boldsymbol{x}$, we denote the ℓ_2-norm as $\|\boldsymbol{x}\|_2 \triangleq \sqrt{\langle \boldsymbol{x}, \boldsymbol{x}\rangle} = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$, where $\langle \boldsymbol{x}, \boldsymbol{x}\rangle = \boldsymbol{x}^\top \boldsymbol{x}$ is the inner product. Lastly, $\mathsf{X} \perp\!\!\!\perp \mathsf{Y}$ denotes independence of X and Y.

Table 1. Nomenclature

Symb.	Description	Ref.
$F(\boldsymbol{x})$	Objective function	Eq. (1)
$f_i(\boldsymbol{x})$	ith component of F	Eq. (1)
$\nabla f_B(\boldsymbol{x})$	Minibatch subsampling estimator of ∇F	Eq. (4)
B	Minibatch of component indices	Eq. (3)
π	Minibatch subsampling strategy	Eq. (3)
b_{eff}	Effective sample size of π	Eq. (5)
$\boldsymbol{g}_i(\boldsymbol{x})$	Unbiased stochastic estimator of ∇f_i	Eq. (7)
$\boldsymbol{g}_i(\boldsymbol{x}; \boldsymbol{\eta})$	Integrand of estimator $\boldsymbol{g}_i(\boldsymbol{x})$	Eq. (7)
$\boldsymbol{g}_B(\boldsymbol{x})$	Doubly stochastic estimator of ∇F	Eq. (8)
$\mathcal{L}_{\text{sub}}$	ER constant (Definition 1) of π	Assu. 5
$\mathcal{L}_i$	ER constant (Definition 1) of $\boldsymbol{g}_i$	Assu. 6
τ^2	BV constant (Definition 2) of π	Assu. 7
σ_i^2	BV constant (Definition 2) of $\boldsymbol{g}_i$	Assu. 7

2.1. Stochastic Gradient Descent on Finite-Sums

Stochastic gradient descent (SGD) is an optimization algorithm that repeats the steps

$$\boldsymbol{x}_{t+1} = \Pi_{\mathcal{X}}\left(\boldsymbol{x}_t - \gamma_t \boldsymbol{g}(\boldsymbol{x}_t)\right),$$

where, $\Pi_{\mathcal{X}}$ is a projection operator onto $\mathcal{X}$, $(\gamma_t)_{i=0}^{T-1}$ is some stepsize schedule, $\boldsymbol{g}(\boldsymbol{x})$ is an unbiased estimate of $\nabla F(\boldsymbol{x})$.

Finite-Sum Problems. When the objective can be represented as a "finite sum" it is typical to approximate the gradients of the objective as

$$\nabla F(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\boldsymbol{x}) = \mathbb{E}_{B\sim\pi}\left[\frac{1}{b}\sum_{i\in B}\nabla f_i(\boldsymbol{x})\right], \quad (3)$$

where $B \sim \pi$ is an index set of cardinality $|B| = b$, or "minibatch," formed by subsampling over the datapoint indices $\{1,\dots,n\}$. More formally, we are approximating ∇F using the (minibatch) subsampling estimator

$$\nabla f_B(\boldsymbol{x}) \triangleq \frac{1}{b}\sum_{i\in B}\nabla f_i(\boldsymbol{x}), \quad (4)$$

where the performance of this estimator, or equivalently, of the subsampling strategy π, can be quantified by its variance

$$\mathrm{tr}\mathbb{V}\left[\nabla f_B(\boldsymbol{x})\right] = \frac{1}{b_{\mathrm{eff}}}\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(\boldsymbol{x}) - \nabla F(\boldsymbol{x})\right\|_2^2}_{\text{(unit) subsampling variance}}, \quad (5)$$

where we say b_{eff} is the "effective sample size" of π. For instance, independent subsampling achieves $b_{\mathrm{eff}} = b$, and sampling without replacement, also known as "b-nice sampling" (Gower et al., 2019; Richtárik & Takáč, 2016; Csiba & Richtárik, 2018), achieves $b_{\mathrm{eff}} = {}^{(n-1)b}/_{n-b}$ (Lemma 2).

2.2. Doubly Stochastic Gradients

For problems where the components are defined as intractable expectations as in Eq. (2), we have to rely on an additional Monte Carlo approximation step such as

$$\nabla F(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\boldsymbol{x}) = \mathbb{E}_{B\sim\pi}\left[\frac{1}{b}\sum_{i\in B}\mathbb{E}_{\boldsymbol{\eta}\sim\varphi}\left[\nabla f_i(\boldsymbol{x};\boldsymbol{\eta})\right]\right]$$
$$= \mathbb{E}_{B\sim\pi,\ \boldsymbol{\eta}_j\sim\varphi}\left[\frac{1}{mb}\sum_{i\in B}\sum_{j=1}^{m}\nabla f_i(\boldsymbol{x};\boldsymbol{\eta}_j)\right], \quad (6)$$

where $\boldsymbol{\eta}_j \sim \varphi$ are m independently and identically distributed (*i.i.d.*) Monte Carlo samples from φ.

Doubly Stochastic Gradient Consider an unbiased estimator of the component gradient ∇f_i such that

$$\mathbb{E}\boldsymbol{g}_i(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\eta}\sim\varphi}\boldsymbol{g}_i(\boldsymbol{x};\boldsymbol{\eta}) = \nabla f_i(\boldsymbol{x}), \quad (7)$$

where $\boldsymbol{g}_i(\boldsymbol{x};\boldsymbol{\eta})$ is the measurable integrand. Using these, we can estimate ∇F through the *doubly stochastic* gradient estimator

$$\boldsymbol{g}_B(\boldsymbol{x}) \triangleq \frac{1}{b}\sum_{i\in B}\boldsymbol{g}_i(\boldsymbol{x}), \quad (8)$$

We separately define the integrand $\boldsymbol{g}(\boldsymbol{x};\boldsymbol{\eta})$ since, in practice, a variety of unbiased estimators of ∇f_i can be obtained by appropriately defining the integrand $\boldsymbol{g}_i$. For example, one can form the m-sample "naive" Monte Carlo estimator by setting

$$\boldsymbol{g}_i(\boldsymbol{x};\boldsymbol{\eta}) = \frac{1}{m}\sum_{j=1}^{m}\nabla f_i(\boldsymbol{x};\boldsymbol{\eta}_j),$$

where $\boldsymbol{\eta} = [\boldsymbol{\eta}_1,\dots,\boldsymbol{\eta}_m] \sim \varphi^{\otimes m}$.

Dependent Component Gradient Estimators. Notice that, in Eq. (6), the subcomponents in the batch share the Monte Carlo samples, which may occur in practice. This means $\boldsymbol{g}_i$ and $\boldsymbol{g}_j$ in the same batch are dependent and, in the worst case, positively correlated, which complicates the analysis. While it is possible to make the estimators independent by sampling m unique Monte Carlo samples for each component (mb Monte Carlo samples in total) as highlighted by Kingma et al. (2015), it is common to use dependent estimators for various practical reasons:

1. **ERM with Randomized Smoothing**: In the ERM context, recent works have studied the generalization benefits of randomly perturbing the model weights before computing the gradient (Orvieto et al., 2023; Liu et al., 2021). When subsampling is used, perturbing the weights independently for each datapoint is computationally inefficient. Therefore, the perturbation is shared across the batch, creating dependence.
2. **Black-Box Variational inference** (Titsias & Lázaro-Gredilla, 2014; Kucukelbir et al., 2017): Here, each component can be decomposed as
$$f_i(\boldsymbol{x};\boldsymbol{\eta}) = \ell_i(\boldsymbol{x};\boldsymbol{\eta}) + r(\boldsymbol{x};\boldsymbol{\eta}),$$
where ℓ_i is the log likelihood and r is the log-density of the prior. By sharing $(\boldsymbol{\eta}_j)_{j=1}^m$, r only needs to be evaluated m times. To create independent estimators, it needs to be evaluated mb times instead, but r can be expensive to compute.
3. **Random feature kernel regression** with doubly SGD (Dai et al., 2014): The features are shared across the batch [1]. This reduces the peak memory requirement from $bmd_{\boldsymbol{\eta}}$, where $d_{\boldsymbol{\eta}}$ is the size of the random features, to $md_{\boldsymbol{\eta}}$.

One of the analysis goals of this work is to characterize the effect of dependence in the context of SGD.

2.3. Technical Assumptions on Gradient Estimators

To establish convergence of SGD, contemporary analyses use the "variance transfer" strategy (Moulines & Bach, 2011; Johnson & Zhang, 2013; Nguyen et al., 2018; Gower et al., 2019; 2021b). That is, by assuming the gradient noise satisfies some condition resembling smoothness, it is possible to bound the gradient noise on some arbitrary point $\boldsymbol{x}$ by the gradient variance on the solution set $\boldsymbol{x}_* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$.

ER Condition. In this work, we will use the *expected residual* (ER) condition by Gower et al. (2021a):

[1] See the implementation at https://github.com/zixu1986/Doubly_Stochastic_Gradients

Definition 1 (Expected Residual; ER). A gradient estimator $\boldsymbol{g}$ of $F : \mathcal{X} \to \mathbb{R}$ is said to satisfy ER $(\mathcal{L})$ if

$$\mathrm{tr}\mathbb{V}\left[\boldsymbol{g}\left(\boldsymbol{x}\right) - \boldsymbol{g}\left(\boldsymbol{x}_*\right)\right] \leq 2\mathcal{L}\left(F(\boldsymbol{x}) - F(\boldsymbol{x}_*)\right),$$

for some $0 < \mathcal{L} < \infty$ and all $\boldsymbol{x} \in \mathcal{X}$ and all $\boldsymbol{x}_* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} F\left(\boldsymbol{x}\right)$.

When f is convex, a weaker form can be used: We will also consider the *convex* variant of the ER condition that uses the Bregman divergence defined as

$$\mathrm{D}_{\phi}\left(\boldsymbol{y}, \boldsymbol{x}\right) \triangleq \phi\left(\boldsymbol{y}\right) - \phi\left(\boldsymbol{x}\right) - \left\langle \nabla\phi\left(\boldsymbol{x}\right), \boldsymbol{y} - \boldsymbol{x}\right\rangle,$$

$\forall(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^2$, where $\phi : \mathcal{X} \to \mathbb{R}$ is a convex function.

Why the ER condition? A way to think about the ER condition is that it corresponds to the "variance form" equivalent of the expected smoothness (ES) condition by Gower et al. (2021b) defined as

$$\mathbb{E}\|\boldsymbol{g}\left(\boldsymbol{x}\right) - \boldsymbol{g}\left(\boldsymbol{x}_*\right)\|_2^2 \leq 2\mathcal{L}\left(F\left(\boldsymbol{x}\right) - F\left(\boldsymbol{x}_*\right)\right), \qquad \text{(ES)}$$

but is slightly weaker, as shown by Gower et al. (2021a). The main advantage of the ER condition is that, due to the properties of the variance, it composes more easily:

Proposition 1. *Let $\boldsymbol{g}$ satisfy* ER $(\mathcal{L})$. *Then, the m-sample i.i.d. average of $\boldsymbol{g}$ satisfy* ER $(\mathcal{L}/m)$.

BV Condition. From the ER property, the gradient variance on any point $\boldsymbol{x} \in \mathcal{X}$ can be bounded by the variance on the solution set as long as the following holds:

Definition 2 (Bounded Gradient Variance). A gradient estimator $\boldsymbol{g}$ of $F : \mathcal{X} \to \mathbb{R}$ satisfies BV $\left(\sigma^2\right)$ if

$$\mathrm{tr}\mathbb{V}\left[\boldsymbol{g}\left(\boldsymbol{x}_*\right)\right] \leq \sigma^2$$

for some $\sigma^2 < \infty$ and all $\boldsymbol{x}_* \in \arg\max_{\boldsymbol{x}\in\mathcal{X}} F\left(\boldsymbol{x}\right)$.

2.4. Convergence Guarantees for SGD

Sufficiency of ER and BV. From the ER and BV conditions, other popular conditions such as ES (Gower et al., 2021b) and ABC (Khaled & Richtárik, 2023) can be established with minimal additional assumptions. As a result, we retrieve the previous convergence results on SGD established for various objective function classes:

- ➤ strongly convex (Gower et al., 2019),
- ➤ quasar convex (+PL) (Gower et al., 2021a),
- ➤ smooth (+PL) (Khaled & Richtárik, 2023).

(Note: quasar convexity is strictly weaker than convexity Guminov et al., 2023; PL: Polyak-Łojasiewicz.) (See also the comprehensive treatment by Garrigos & Gower, 2023.) Therefore, ER and BV are sufficient conditions for SGD to converge on problem classes typically considered in SGD convergence analysis.

In this work, we will specifically focus on smooth and strongly convex objectives:

Assumption 1. There exists some μ, L satisfying $0 < \mu \leq L < \infty$ suc that the objective function $F : \mathcal{X} \to \mathbb{R}$ is μ-strongly convex and L-smooth as

$$F\left(\boldsymbol{y}\right) - F\left(\boldsymbol{x}\right) \geq \left\langle \nabla F\left(\boldsymbol{x}\right), \boldsymbol{y} - \boldsymbol{x}\right\rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

$$F\left(\boldsymbol{y}\right) - F\left(\boldsymbol{x}\right) \leq \left\langle \nabla F\left(\boldsymbol{x}\right), \boldsymbol{y} - \boldsymbol{x}\right\rangle + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

hold for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^2$.

Also, we will occasionally assume that F is comprised of a finite sum of convex and smooth components:

Assumption 2. The objective function $F : \mathcal{X} \to \mathbb{R}$ is a finite sum as $F = \frac{1}{n}\left(f_1 + \ldots + f_n\right)$, where each component is L_i-smooth and convex such that

$$\|\nabla f_i\left(\boldsymbol{x}\right) - \nabla f_i\left(\boldsymbol{y}\right)\|_2^2 \leq 2L_i\,\mathrm{D}_{f_i}\left(\boldsymbol{x}, \boldsymbol{y}\right)$$

holds for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^2$.

Note that Assumption 2 alone already implies that F is convex and $L_{\max}$-smooth with $L_{\max} = \max\left\{L_1, \ldots, L_n\right\}$.

Why focus on strongly convex functions? We focus on strongly convex objectives as the effect of stochasticity is the most detrimental: in the deterministic setting, one only needs $\mathcal{O}\left(\log\left(1/\epsilon\right)\right)$ iterations to achieve an ϵ-accurate solution. But with SGD, one actually needs $\mathcal{O}\left(1/\epsilon\right)$ iterations due to noise. As such, we can observe a clear contrast between the effect of optimization and noise in this setting.

With that said, for completeness, we provide full proof of convergence on strongly convex-smooth objectives:

Lemma 1. *Let the objective F satisfy Assumption 1 and the gradient estimator $\boldsymbol{g}$ satisfy* ER $(\mathcal{L})$ *and* BV $\left(\sigma^2\right)$. *Then, the last iterate of SGD is ϵ-close to the global optimum $\boldsymbol{x}_* = \arg\min_{\boldsymbol{x}\in\mathcal{X}} F\left(\boldsymbol{x}\right)$ such that $\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}_*\|_2^2 \leq \epsilon$ after a number of iterations at least*

$$T \geq 2\max\left(\frac{\sigma^2}{\mu^2}\frac{1}{\epsilon}, \frac{\mathcal{L} + L}{\mu}\right)\log\left(2\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2\frac{1}{\epsilon}\right)$$

and the fixed stepsize

$$\gamma = \min\left(\frac{\epsilon\mu}{2\sigma^2}, \frac{1}{2\left(\mathcal{L} + L\right)}\right).$$

See the *full proof* in page 22.

Note that our complexity guarantee is only $\mathcal{O}(1/\epsilon\log\left(1/\epsilon\right))$ due to the use of a fixed stepsize. It is also possible to establish a $\mathcal{O}(1/\epsilon)$ guarantee using decreasing stepsize schedules proposed by Gower et al. (2019); Stich (2019). In practice, these schedules are rarely used, and the resulting complexity guarantees are less clear than with fixed stepsizes. Therefore, we will stay on fixed stepsizes.

3. Main Results

3.1. Doubly Stochastic Gradients

Table 2. Rosetta Stone

§3.1.1		§3.1.2
$\boldsymbol{x}_i$	$\leftrightarrow$	$\boldsymbol{g}_i$
$\boldsymbol{x}_B$	$\leftrightarrow$	$\boldsymbol{g}_B$
$\bar{\boldsymbol{x}}_i$	$\leftrightarrow$	∇f_i
$\bar{\boldsymbol{x}}$	$\leftrightarrow$	∇F

First, while taming notational complexity, we will prove a general result that holds for combining unbiased but potentially correlated estimators through subsampling. All of the later results on SGD will fall out as special cases following the correspondence in Table 2.

3.1.1. General Variance Bound

Theoretical Setup. Consider the problem of estimating the population mean $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \bar{\boldsymbol{x}}_i$ with a collection of RVs $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, each an unbiased estimator of the component $\bar{\boldsymbol{x}}_i = \mathbb{E}\boldsymbol{x}_i$. Then, any subsampled ensemble

$$\boldsymbol{x}_B \triangleq \frac{1}{b}\sum_{i\in B} \boldsymbol{x}_i \quad \text{with} \quad B \sim \pi,$$

where π is an unbiased subsampling strategy with an effective sample size of b_{eff}, is also an unbiased estimator of $\bar{\boldsymbol{x}}$. The goal is to analyze how the variance of the component estimators $\operatorname{tr}\mathbb{V}\boldsymbol{x}_i$ for $i = 1, \dots, n$ and the variance of π affect the variance of $\boldsymbol{x}_B$. The following condition characterizes the correlation between the component estimators:

Assumption 3. The component estimators $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ have finite variance $\operatorname{tr}\mathbb{V}\boldsymbol{x}_i < \infty$ for all $i = 1, \dots, n$ and, there exists some $\rho \in [0, 1]$ for all $i \neq j$ such that

$$\operatorname{tr}\operatorname{Cov}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \leq \rho\sqrt{\operatorname{tr}\mathbb{V}\boldsymbol{x}_i}\sqrt{\operatorname{tr}\mathbb{V}\boldsymbol{x}_j}.$$

Remark 1. Assumption 3 always holds with $\rho = 1$ as a basic consequence of the Cauchy-Schwarz inequality.

Remark 2. For a collection of mutually independent estimators $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ such that $\boldsymbol{x}_i \perp\!\!\!\perp \boldsymbol{x}_j$ for all $i \neq j$, Assumption 3 holds with $\rho = 0$.

Remark 3. The equality in Assumption 3 holds with $\rho = 0$ for independent estimators, while it holds with $\rho = 1$ when they are perfectly positively correlated such that, for all $i \neq j$, there exists some constant $\alpha_{ij} \geq 0$ such that $\boldsymbol{x}_i = \alpha_{i,j}\boldsymbol{x}_j$

Theorem 1. *Let the component estimators $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ satisfy Assumption 3. Then, the variance of the doubly stochastic estimator $\boldsymbol{x}_B$ is bounded as*

$$\operatorname{tr}\mathbb{V}\left[\boldsymbol{x}_B\right] \leq V_{\text{com}} + V_{\text{cor}} + V_{\text{sub}},$$

where

$$V_{\text{com}} = \left(\frac{\rho}{b_{\text{eff}}} + \frac{1-\rho}{b}\right)\left(\frac{1}{n}\sum_{i=1}^n \operatorname{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]\right),$$

$$V_{\text{cor}} = \rho\left(1 - \frac{1}{b_{\text{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^n \sqrt{\operatorname{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]}\right)^2, \text{ and}$$

$$V_{\text{sub}} = \frac{1}{b_{\text{eff}}}\frac{1}{n}\sum_{i=1}^n \|\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}\|_2^2.$$

Equality holds when the equality in Assumption 3 holds.

Proof. We start from the law of total (co)variance,

$$\operatorname{tr}\mathbb{V}\left[\boldsymbol{x}_B\right] = \underbrace{\mathbb{E}_\pi\left[\operatorname{tr}\mathbb{V}\left[\boldsymbol{x}_B \mid B\right]\right]}_{\text{Variance of ensemble}} + \underbrace{\operatorname{tr}\mathbb{V}_\pi\left[\mathbb{E}\left[\boldsymbol{x}_B \mid B\right]\right]}_{\text{Variance of subsampling}}.$$

This splits the variance into the variance of the specific ensemble of B and subsampling variance. The main challenge is to relate the variance of the ensemble of B with the variance of the individual estimators in the sum

$$\mathbb{E}_\pi\left[\operatorname{tr}\mathbb{V}\left[\boldsymbol{x}_B \mid B\right]\right] = \mathbb{E}_\pi\left[\operatorname{tr}\mathbb{V}\left[\frac{1}{b}\sum_{i\in B}\boldsymbol{x}_i\right]\right]. \tag{9}$$

Since the individual estimators may not be independent, analyzing the variance of the sum can be tricky. However, the following lemma holds generally:

Lemma 9. *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_b$ be a collection of vector-variate RVs dependent on some random variable B satisfying Assumption 3. Then, the expected variance of the sum of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_b$ conditioned on B is bounded as*

$$\mathbb{E}\left[\operatorname{tr}\mathbb{V}\left[\sum_{i=1}^b \boldsymbol{x}_i \mid B\right]\right] \leq \rho\mathbb{V}\left[S\right] + \rho(\mathbb{E}S)^2 + (1-\rho)\,\mathbb{E}\left[V\right],$$

where

$$S = \sum_{i=1}^b \sqrt{\operatorname{tr}\mathbb{V}\left[\boldsymbol{x}_i \mid B\right]} \text{ and } V = \sum_{i=1}^b \operatorname{tr}\mathbb{V}\left[\boldsymbol{x}_i \mid B\right].$$

Equality holds when the equality in Assumption 3 holds.

Here, S is the sum of conditional standard deviations, while V is the sum of conditional variances. Notice that the "variance of the variances" is playing a role: if we reduce the subsampling variance, then the variance of the ensemble, V_{com}, also decreases.

The rest of the proof, along with the proof of Lemma 9, can be found in Appendix B.3 page 23. □

In Theorem 1, V_{com} is the contribution of the variance of the component estimators, while V_{cor} is the contribution of the correlation between component estimators , and V_{sub} is the subsampling variance.

Monte Carlo with Subampling Without Replacement. Theorem 1 is very general: it encompasses both the correlated and uncorrelated cases and matches the constants of all of the important special cases. We will demonstrate this in the following corollary along with variance reduction by Monte Carlo averaging of m i.i.d. samples. That is, we subsample over $\boldsymbol{x}_1^m, \dots, \boldsymbol{x}_n^m$, where each estimator is an m-sample Monte Carlo estimator:

$$\boldsymbol{x}_i^m \triangleq \frac{1}{m}\sum_{j=1}^m \boldsymbol{x}_i^{(j)},$$

where $\boldsymbol{x}_i^{(1)}, \dots, \boldsymbol{x}_i^{(m)}$ are i.i.d replications with mean $\bar{\boldsymbol{x}}_i = \mathbb{E}\boldsymbol{x}_i^{(j)}$. Then, the variance of the doubly stochastic estimator $\boldsymbol{x}_B$ of the mean $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \bar{\boldsymbol{x}}_i$ defined as

$$\boldsymbol{x}_B^m \triangleq \frac{1}{b}\sum_{i\in B}\boldsymbol{x}_i^m \quad \text{with} \quad B \sim \pi,$$

can be bounded as follows:

Corollary 1. *For each $j = 1, \dots, m$, let $\boldsymbol{x}_1^{(j)}, \dots, \boldsymbol{x}_n^{(j)}$ satisfy Assumption 3. Then, the variance of the doubly stochastic estimator $\boldsymbol{x}_B^m$ of the mean $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \bar{\boldsymbol{x}}_i$, where π is b-minibatch sampling without replacement, satisfy the following corollaries:*

(i) *$\rho = 1$ and $1 < b < n$:*

$$\operatorname{tr}\mathbb{V}\left[\boldsymbol{x}_B^m\right] \leq \frac{n-b}{(n-1)mb}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i^2\right) + \frac{n(b-1)}{(n-1)mb}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i\right)^2 + \frac{n-b}{(n-1)b}\tau^2$$

(ii) *$\rho = 1$ and $b = 1$:*

$$\operatorname{tr}\mathbb{V}\left[\boldsymbol{x}_B^m\right] \leq \frac{1}{m}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i^2\right) + \tau^2$$

(iii) *$\rho = 1$ and $b = n$:*

$$\operatorname{tr}\mathbb{V}\left[\boldsymbol{x}_B^m\right] \leq \frac{1}{m}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i\right)^2$$

(iv) *$\sigma_i = 0$ for all $i = 1, \dots, n$:*

$$\operatorname{tr}\mathbb{V}\left[\boldsymbol{x}_B^m\right] \leq \frac{n-b}{(n-1)b}\tau^2,$$

(v) *$\rho = 0$:*

$$\operatorname{tr}\mathbb{V}\left[\boldsymbol{x}_B^m\right] \leq \frac{1}{mb}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i^2\right) + \frac{n-b}{(n-1)b}\tau^2$$

where, for all $i = 1, \dots, n$ and any $j = 1, \dots, m$,

$\sigma_i^2 = \operatorname{tr}\mathbb{V}\,\boldsymbol{x}_i^{(j)}$ *is invidual variance and*

$\tau^2 = \frac{1}{n}\sum_{i=1}^n \|\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}\|_2^2$ *is the subsampling variance.*

Remark 4 (For dependent estimators, increasing b also reduces component variance.). Notice that, for case of $\rho = 1$, Corollary 1 (i), the term with $\frac{1}{n}\sum_{i=1}^n \sigma_i^2$ is reduced in a rate of $\mathcal{O}(1/mb)$. This means reducing the subsampling noise by increasing b also reduces the noise of estimating each component. Furthermore, the first term dominates the second term as

$$\left(\frac{1}{n}\sum_{i=1}^n \sigma_i\right)^2 \leq \frac{1}{n}\sum_{i=1}^n \sigma_i^2,$$

as stated by Jensen's inequality. Therefore, despite correlations, increasing b will have a more significant effect since it reduces both dominant terms $\frac{1}{n}\sum_{i=1}^n \sigma_i^2$ and τ^2.

Remark 5. When independent estimators are used, Corollary 1 (v) shows that increasing b reduces the full variance in a $\mathcal{O}(1/b)$ rate, but increasing m does not.

Remark 6. Corollary 1 achieves all known endpoints in the context of SGD: For $b = n$ (full batch), doubly SGD reduces to SGD with a Monte Carlo estimator, where there is no subsampling noise (no τ^2). When the Monte Carlo noise is 0, then doubly SGD reduces to SGD with a subsampling estimator (no σ_i), retrieving the result of Gower et al. (2019).

3.1.2. Gradient Variance Conditions for SGD

From Theorem 1, we can establish the ER and BV conditions (Section 2.3) of the doubly stochastic gradient estimators. Following the notation in Section 2.2, we will denote the doubly stochastic gradient estimator as $\boldsymbol{g}_B$, which combines the estimators $\boldsymbol{g}_1, \dots, \boldsymbol{g}_n$ according to the subsampling strategy $B \sim \pi$, which achieves an effective sample size of b_{eff}. We will also use the corresponding minibatch subsampling estimator ∇f_B for the analysis.

Assumption 4. For all $\boldsymbol{x} \in \mathcal{X}$, the component gradient estimators $\boldsymbol{g}_1(\boldsymbol{x}), \dots, \boldsymbol{g}_n(\boldsymbol{x})$ satisfy Assumption 3 with some $\rho \in [0, 1]$.

Again, this assumption is always met with $\rho = 1$ and holds with $\rho = 0$ if the estimators are independent.

Assumption 5. The subsampling estimator ∇f_B satisfies the ER ($\mathcal{L}_{\text{sub}}$) condition in Definition 1.

This is a classical assumption used to analyze SGD on finite sums and is automatically satisfied by Assumption 2. (See Lemma 10 in Appendix B.4.3 for a proof.)

Assumption 6. For all $i = 1, \dots, n$ and $\boldsymbol{x} \in \mathcal{X}$ and global minimizers $\boldsymbol{x}_* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x}_*)$, the component gradient estimator $\boldsymbol{g}_i$ satisfies at least one of the following variants of the ER condition:

(A^{CVX}) $\operatorname{tr}\mathbb{V}\left[\boldsymbol{g}_i(\boldsymbol{x}) - \boldsymbol{g}_i(\boldsymbol{y})\right] \leq 2\mathcal{L}_i \mathrm{D}_{f_i}(\boldsymbol{x}, \boldsymbol{y})$, where f_i is convex.

(A^{ITP}) $\operatorname{tr}\mathbb{V}\left[\boldsymbol{g}_i(\boldsymbol{x}) - \boldsymbol{g}_i(\boldsymbol{y})\right] \leq 2\mathcal{L}_i\left(f_i(\boldsymbol{x}) - f_i(\boldsymbol{x}_*)\right)$ where $f_i(\boldsymbol{x}) \geq f_i(\boldsymbol{x}_*)$.

(B) $\operatorname{tr}\mathbb{V}\left[\boldsymbol{g}_i(\boldsymbol{x}) - \boldsymbol{g}_i(\boldsymbol{y})\right] \leq 2\mathcal{L}_i\left(F(\boldsymbol{x}) - F(\boldsymbol{x}_*)\right)$.

Each of these assumptions holds under different assumptions and problem setups. For instance, A^{CVX} holds only under componentwise convexity, while A^{ITP} requires majorization $f_i(\boldsymbol{x}) \geq f_i(\boldsymbol{x}_*)$, which is essentially assuming "interpolation" (Vaswani et al., 2019; Ma et al., 2018; Gower et al., 2021a) in the ERM context. Among these, (B) is the strongest since it directly relates the individual components $f_1, \dots, f_n$ with the full objective F.

We now state our result establishing the ER condition:

Theorem 2. *Let Assumption 4 to 6 hold. Then, we have:*

(i) *If* (A^{CVX}) *or* (A^{ITP}) *hold, $\boldsymbol{g}_B$ satisfies* ER ($\mathcal{L}_{\text{A}}$).

(ii) *If* (B) *holds, $\boldsymbol{g}_B$ satisfies* ER ($\mathcal{L}_{\text{B}}$).

where $\mathcal{L}_{\max} = \max\{\mathcal{L}_1, \dots, \mathcal{L}_n\}$,

$$\mathcal{L}_{\text{A}} = \left(\frac{\rho}{b_{\text{eff}}} + \frac{1-\rho}{b}\right)\mathcal{L}_{\max} + \rho\left(1 - \frac{1}{b_{\text{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^n \mathcal{L}_i\right) + \frac{\mathcal{L}_{\text{sub}}}{b_{\text{eff}}}$$

$$\mathcal{L}_{\mathrm{B}} = \left(\frac{\rho}{b_{\mathrm{eff}}} + \frac{1-\rho}{b}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i\right) + \rho\left(1 - \frac{1}{b_{\mathrm{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\mathcal{L}_i}\right)^2 + \frac{\mathcal{L}_{\mathrm{sub}}}{b_{\mathrm{eff}}}.$$

See the *full proof* in page 25.

Remark 7. Assuming the conditions in Assumption 6 hold with the same value of $\mathcal{L}_i$, the inequality

$$\left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\mathcal{L}_i}\right)^2 \leq \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i \leq \mathcal{L}_{\max},$$

implies $\mathcal{L}_{\mathrm{B}} \leq \mathcal{L}_{\mathrm{A}}$.

Meanwhile, The BV condition follows by assuming equivalent conditions on each component estimator:

Assumption 7. Variance is bounded for all $\boldsymbol{x}_* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$ such that the following hold:

1. $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\boldsymbol{x}_*)\|_2^2 \leq \tau^2$ for some $\tau^2 < \infty$ and,
2. $\mathrm{tr}\mathbb{V}[\boldsymbol{g}_i(\boldsymbol{x}_*)] \leq \sigma_i^2$ for some $\sigma_i^2 < \infty$, for all $i = 1, \dots, n$.

Based on these, Theorem 1 immediately yields the result:

Theorem 3. *Let Assumption 4 and 7 hold. Then, $\boldsymbol{g}_B$ satisfies* BV(σ^2)*, where*

$$\sigma^2 = \left(\frac{\rho}{b_{\mathrm{eff}}} + \frac{1-\rho}{b}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\right) + \rho\left(1 - \frac{1}{b_{\mathrm{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i\right)^2 + \frac{\tau^2}{b_{\mathrm{eff}}}.$$

Equality in Definition 2 holds if equality in Assumption 4 holds.

See the *full proof* in page 27.

As discussed in Section 2.4, Theorems 2 and 3 are sufficient to guarantee convergence of doubly SGD. For completeness, let us state a specific result for $\rho = 1$:

Corollary 2. *Let the objective F satisfy Assumption 1 and 2, the global optimum $\boldsymbol{x}_* = \arg\min_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$ be a stationary point of F, the component gradient estimators $\boldsymbol{g}_1, \dots, \boldsymbol{g}_n$ satisfy Assumption 6 (B) and 7, and π be b-minibatch sampling without replacement. Then the last iterate of SGD with $\boldsymbol{g}_B$ is ϵ-close to $\boldsymbol{x}_*$ as $\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}_*\|_2^2 \leq \epsilon$ after a number of iterations of at least*

$$T \geq 2\max\left(C_{\mathrm{var}}\frac{1}{\epsilon},\ C_{\mathrm{bias}}\right)\log\left(2\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2\frac{1}{\epsilon}\right)$$

for some fixed stepsize where

$$C_{\mathrm{var}} = \frac{2}{b}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\sigma_i^2}{\mu^2}\right) + 2\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\sigma_i}{\mu}\right)^2 + \frac{2}{b}\frac{\tau^2}{\mu^2},$$

$$C_{\mathrm{bias}} = \frac{2}{b}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\mathcal{L}_i}{\mu}\right) + 2\left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{\mathcal{L}_i}{\mu}}\right)^2 + \frac{2}{b}\frac{L}{\mu}.$$

See the *full proof* in page 28.

3.2. Random Reshuffling of Stochastic Gradients

We now move to our analysis of SGD with random reshuffling (SGD-RR). In the doubly stochastic setting, this corresponds to reshuffling over stochastic estimators instead of gradients, which we will denote as doubly SGD-RR. In practice, doubly SGD-RR is often observed to converge faster than doubly SGD, even when dependent estimators are used.

3.2.1. Algorithm

Doubly SGD-RR The algorithm is stated as follows:

❶ Reshuffle and partition the gradient estimators into minibatches of size b as $P = \{P_1, \dots, P_p\}$, where $p = n/b$ is the number of partitions or minibatches.

❷ Perform gradient descent for $i = 1, \dots, p$ steps as

$$\boldsymbol{x}_k^{i+1} = \Pi_{\mathcal{X}}\left(\boldsymbol{x}_k^i - \gamma\boldsymbol{g}_{P_i}(\boldsymbol{x}_k^i)\right)$$

❸ $k \leftarrow k+1$ and go back to step ❶.

(We assume n is an integer multiple of b for clarity.) Here, $i = 1, \dots, p$ denotes the step within the epoch, $k = 1, \dots, K$ denotes the epoch number.

3.2.2. Proof Sketch

Why SGD-RR is Faster A key aspect of random reshuffling in the finite sum setting (SGD-RR) is that it uses conditionally biased gradient estimates. Because of this, on strongly convex finite sums, Mishchenko et al. (2020) show that the Lyapunov function for random reshuffling is not the usual $\|\boldsymbol{x}_k^i - \boldsymbol{x}_*\|_2^2$, but some *biased* Lyapunov function $\|\boldsymbol{x}_k^i - \boldsymbol{x}_*^k\|_2^2$, where the reference point is

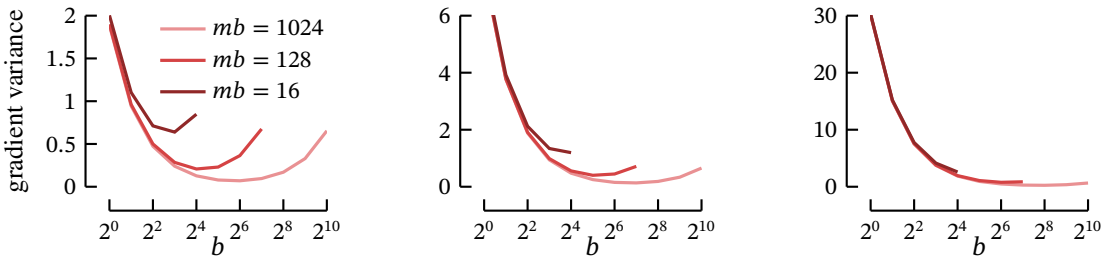
$$\boldsymbol{x}_*^i \triangleq \Pi_{\mathcal{X}}\left(\boldsymbol{x}_* - \gamma\sum_{j=0}^{i-1}\nabla f_{P_i}(\boldsymbol{x}_*)\right). \tag{10}$$

Under this definition, the convergence rate of SGD is not determined by the gradient variance anymore; it is determined by the squared error of the Lyapunov reference point, $\|\boldsymbol{x}_*^i - \boldsymbol{x}_*\|_2^2$. There are two key properties of this quantity:

- The peak mean-squared error decreases at a rate of γ^2 with respect to the stepsize γ.
- The squared error is 0 at the following two endpoints: beginning of the epoch and at the end of the epoch.

For some stepsize achieving a $\mathcal{O}(1/T)$ rate on SGD, these two properties combined result in SGD-RR attaining a $\mathcal{O}(1/T^2)$ rate at exactly the end of each epoch.

Is doubly SGD-RR as Fast as SGD-RR? Unfortunately, doubly SGD-RR does not achieve the same rate as SGD-RR. Since stochastic gradients are used in addition to reshuffling, doubly SGD-RR deviates from the path that minimizes the biased Lyapunov function. Still, doubly SGD-RR does have provable benefits.



(a) Low Hetero. ($s = 1$) (b) Mid. Hetero. ($s = 2$) (c) High Hetero. ($s = 4$)

Figure 1. **Trade-off between b and m on the gradient variance $\mathrm{tr}\mathbb{V}\boldsymbol{g}(\boldsymbol{x}_*)$ under varying budgets $m \times b$.** The problem is a finite sum of $d = 10$, $n = 1024$ isotropic quadratics with smoothness constants sampled as $L_i \sim \text{Inv-Gamma}(1/2, 1/2)$ and stationary points sampled as $\boldsymbol{x}_i^* \sim \mathcal{N}\left(\mathbf{0}_d, s^2\mathbf{I}_d\right)$, where the gradient has additive noise of $\boldsymbol{\eta} \sim \mathcal{N}\left(\mathbf{0}_d, \mathbf{I}_d\right)$. Larger s means more heterogeneous data.

3.2.3. Complexity Analysis

We provide the general complexity guarantee for doubly SGD-RR on strongly convex objectives with μ-strongly convex components and fully correlated component estimators ($\rho = 1$):

Theorem 4. *Let the objective F satisfy Assumption 1 and 2, where each component f_i is additionally μ-strongly convex, and Assumption 6 ($\mathrm{A}^{\mathrm{CVX}}$), 7 hold. Then, the last iterate $\boldsymbol{x}_T$ of doubly SGD-RR is ϵ-close to the global optimum $\boldsymbol{x}_* = \arg\max_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$ such that $\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}_*\|_2^2 \leq \epsilon$ after a number of iterations of at least*

$$T \geq \max\left(4C_{\mathrm{var}}^{\mathrm{com}}\frac{1}{\epsilon} + C_{\mathrm{var}}^{\mathrm{sub}}\frac{1}{\sqrt{\epsilon}},\ C_{\mathrm{bias}}\right)\log\left(2\left\|\boldsymbol{x}_1^0 - \boldsymbol{x}_*\right\|_2^2\frac{1}{\epsilon}\right)$$

for some fixed stepsize, where $T = Kp = Kn/b$,

$$C_{\mathrm{bias}} = \left(\mathcal{L}_{\max} + L\right)/\mu$$

$$C_{\mathrm{var}}^{\mathrm{com}} = \frac{2}{b}\left(\frac{1}{n}\sum_{i=1}^n\frac{\sigma_i^2}{\mu^2}\right) + 2\left(\frac{1}{n}\sum_{i=1}^n\frac{\sigma_i}{\mu}\right)^2,$$

$$C_{\mathrm{var}}^{\mathrm{sub}} = \sqrt{\frac{L_{\max}}{\mu}}\frac{\sqrt{n}}{b}\frac{\tau}{\mu}.$$

See the *full proof* in page 34.

Remark 8. When $\sigma_i = 0$ for all $i = 1, \ldots, n$, the anytime convergence bound Theorem 5 in the Appendix reduces exactly to Theorem 1 of Mishchenko et al. (2020). Therefore, Theorem 5 is a strict generalization of SGD-RR to the doubly stochastic setting.

Using m-sample Monte Carlo improves the constants as follows:

Corollary 3. *Let the assumptions of Theorem 4 hold. Then, for $1 < b < n$ and m-sample Monte Carlo, the same guarantees hold with the constant*

$$C_{\mathrm{var}}^{\mathrm{com}} = \frac{2}{mb}\left(\frac{1}{n}\sum_{i=1}^n\frac{\sigma_i^2}{\mu^2}\right) + \frac{2}{m}\left(\frac{1}{n}\sum_{i=1}^n\frac{\sigma_i}{\mu}\right)^2.$$

Remark 9. Compared to doubly SGD, doubly SGD-RR improves the dependence on the subsampling noise τ^2 from $\mathcal{O}(1/\epsilon)$ to $\mathcal{O}(1/\sqrt{\epsilon})$. Therefore, random reshuffling does improve the complexity of doubly SGD. Unfortunately, it also means that it does not achieve a better asymptotic complexity as in the finite sum setting. However, non-asymptotically, if the subsampling noise dominates component estimation noise, doubly SGD-RR will behave closely to an $\mathcal{O}(1/\sqrt{\epsilon})$ (or equivalently, $\mathcal{O}(1/T)$) algorithm.

Remark 10. As was the case with independent subsampling, increasing b also reduces component estimation noise for RR-SGD. However, the impact on the complexity is more subtle. Consider that the iteration complexity is

$$\mathcal{O}\left(\kappa_\sigma^2\left(\frac{1}{mb} + \frac{1}{m}\right)\frac{1}{\epsilon} + \kappa\,\kappa_\tau\,\frac{\sqrt{n}}{b}\frac{1}{\sqrt{\epsilon}}\right), \tag{11}$$

where $\kappa_\sigma = \max_{i=1,\ldots,n}\sigma_i/\mu$, $\kappa_\tau = \tau/\mu$ and $\kappa = \max_{i=1,\ldots,n}L_i/\mu$. The $1/\epsilon$ term decreases the fastest with m. Therefore, it might seem that increasing m is advantageous. However, the $1/\sqrt{\epsilon}$ term has a $\mathcal{O}\left(\sqrt{n}\right)$ dependence on the dataset size, which would be non-negligible for large datasets. As a result, in the large n, large ϵ regime, increasing b over m should be more effective.

Remark 11. Eq. (11) also implies that, for dependent estimators, doubly SGD-RR achieves an asymptotic speedup of n/b compared to full-batch SGD with only component estimation noise. Assume that the sample complexity of a single estimate is $\Theta(mb)$ ($\Theta(mn)$ for full-batch). Then, the sample complexity of doubly SGD-RR is $\mathcal{O}(b1/\epsilon)$ and $\mathcal{O}(n1/\epsilon)$ for full-batch SGD. However, the n/b seed-up comes from correlations. Therefore, for independent estimators, the asymptotic complexity of the two is equal.

4. Simulation

Setup We evaluate the insight on the tradeoff between b and m for correlated estimators on a synthetic problem. In particular, we set

$$f_i(\boldsymbol{x};\boldsymbol{\eta}) = \frac{L_i}{2}\left\|\boldsymbol{x} - \boldsymbol{x}_i^* + \boldsymbol{\eta}\right\|_2^2,$$

where the smoothness constants $L_i \sim \text{Inv-Gamma}(1/2, 1/2)$ and the stationary points $\boldsymbol{x}_i^* \sim \mathcal{N}(\mathbf{0}_d, s^2\mathbf{I}_d)$ are sampled randomly, where $\mathbf{0}_d$ is a vector of d zeros and $\mathbf{I}_d$ is a $d \times d$ identity matrix. Then, we compute the gradient variance on the global optimum, corresponding to computing the BV (Definition 2) constant. Note that s^2 here corresponds to the "heterogeneity" of the data. We make the estimators dependent by sharing $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_m$ across the batch.

Results The results are shown in Fig. 1. At low heterogeneity, there exists a "sweet spot" between m and b. However, this sweet spot moves towards large values of b, where, at high heterogeneity levels, the largest values of b are more favorable. Especially in the low budget regime where $mb \ll n$, the largest b values appear to achieve the lowest variance. This confirms our theoretical results that a large b should be preferred on challenging (large number of datapoints, high heterogeneity) problems.

5. Discussions

5.1. Applications

In Appendix C, we establish Assumption 6 and 7 on the following applications:

- **ERM with Randomized Smoothing**: In this problem, we consider ERM, where the model weights are perturbed by noise. This variant of ERM has recently gathered interest as it is believed to improve generalization performance (Orvieto et al., 2023; Liu et al., 2021). In Appendix C.1, we establish Assumption 6 (A^{ITP}) under the interpolation assumption.
- **Reparameterization Gradient**: In certain applications, *e.g.*, variational inference, generative modeling, and reinforcement learning (see Mohamed et al., 2020, §5), the optimization problem is over the parameters of some distribution, which is taken expectation over. Among gradient estimators for this problem, the reparameterization gradient is widely used due to lower variance (Xu et al., 2019). For this, in Appendix C.2, we establish Assumption 6 (A^{CVX}) and (B) by assuming a convexity and smooth integrand.

5.2. Related Works

Unlike SGD in the finite sum setting, doubly SGD has received little interest. Previously, Bietti & Mairal (2017); Zheng & Kwok (2018); Kulunchakov & Mairal (2020) have studied the convergence of variance-reduced gradients (Gower et al., 2020) specific to the doubly stochastic setting under the uniform Lipchitz integrand assumption ($\boldsymbol{g}_i(\cdot;\boldsymbol{\eta})$ is L-Lipschitz for all $\boldsymbol{\eta}$). Although this assumption has often been used in the stochastic optimization literature (Nemirovski et al., 2009; Moulines & Bach, 2011; Shalev-Shwartz et al., 2009; Nguyen et al., 2018), it is easily shown to be restrictive: for some L-smooth $\widehat{f}_i(\boldsymbol{x})$, $\nabla f_i(\boldsymbol{x};\boldsymbol{\eta}) = \nabla \widehat{f}_i(\boldsymbol{x}) + x_1\boldsymbol{\eta}$ is not L-Lipschitz unless the support of $\boldsymbol{\eta}$ is compact. In contrast, we established results under weaker conditions. We also provide a discussion on the relationships of different conditions in Appendix A.

Furthermore, we extended doubly SGD to the case where random reshuffling is used in place of sampling independent batches. In the finite-sum setting, the fact that SGD-RR converges faster than independent subsampling (SGD) has been empirically known for a long time (Bottou, 2009). While Gürbüzbalaban et al. (2021) first demonstrated that SGD-RR can be fast for quadratics, a proof under general conditions was demonstrated recently (Haochen & Sra, 2019): In the strongly convex setting, Mishchenko et al. (2020) Ahn et al. (2020); Nguyen et al. (2021) establish a $\mathcal{O}\left(1/\sqrt{\epsilon}\right)$ complexity to be ϵ-accurate, which is tight in terms of asymptotic complexity (Safran & Shamir, 2020; Cha et al., 2023; Safran & Shamir, 2021).

Lastly, Dai et al. (2014); Xie et al. (2015); Shi et al. (2021) provided convergence guarantees for doubly SGD for ERM of random feature kernel machines. However, these analyses are based on concentration arguments that doubly SGD does not deviate too much from the optimization path of finite-sum SGD. Unfortunately, concentration arguments require stronger assumptions on the noise, and their analysis is application-specific. In contrast, we provide a general analysis under the general ER assumption.

5.3. Conclusions

In this work, we analyzed the convergence of SGD with doubly stochastic and dependent gradient estimators. In particular, we showed that if the gradient estimator of each component satisfies the ER and BV conditions, the doubly stochastic estimator also satisfies both conditions; this implies convergence of doubly SGD.

Practical Recommendations An unusual conclusion of our analysis is that when Monte Carlo is used with minibatch subsampling, it is generally more beneficial to increase the minibatch size b instead of the number of Monte Carlo samples m. That is, for both SGD and SGD-RR, increasing b decreases the variance in a rate close to $1/b$ when **(i)** the gradient variance of the component gradient estimators varies greatly such that $\left(\frac{1}{n}\sum_{i=1}^n \sigma_i\right)^2 \ll \frac{1}{n}\sum_{i=1}^n \sigma_i^2$ or when **(ii)** the estimators are independent as $\rho = 0$. Surprisingly, such a benefit persists even in the interpolation regime $\tau^2 = 0$. On the contrary, when the estimators are both dependent *and* have similar variance, it is necessary to increase both m and b, where a sweet spot between the two exists. However, such a regime is unlikely to occur in practice; in statistics and machine learning applications, the variance of the gradient estimators tends to vary greatly due to the heterogeneity of data.

Acknowledgements

The authors would like to thank the anonymous reviewers for their comments and Jason Altschuler (UPenn) for numerous suggestions that strengthened the work.

K. Kim was supported by a gift from AWS AI to Penn Engineering's ASSET Center for Trustworthy AI; Y.-A. Ma was funded by the NSF Grants [NSF-SCALE MoDL-2134209] and [NSF-CCF-2112665 (TILOS)], the U.S. Department Of Energy, Office of Science, as well as the DARPA AIE program; J. R. Gardner was supported by NSF award [IIS-2145644].

Impact Statement

This paper presents a theoretical analysis of stochastic gradient descent under doubly stochastic noise to broaden our understanding of the algorithm. The work itself is theoretical, and we do not expect direct societal consequences, but SGD with doubly stochastic gradients is widely used in various aspects of machine learning and statistics. Therefore, we inherit the societal impact of the downstream applications of SGD.

References

Ahn, K., Yun, C., and Sra, S. SGD with shuffling: Optimal rates without component convexity and large epoch requirements. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17526–17535. Curran Associates, Inc., 2020. (page 9)

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the IEEE Annual Symposium on Foundations of Computer Science*, FOCS '14, pp. 464–473, USA, October 2014. IEEE Computer Society. (page 1)

Bietti, A. and Mairal, J. Stochastic optimization with variance reduction for infinite datasets with finite sum structure. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1623–1633. Curran Associates, Inc., 2017. (pages 1, 9)

Bottou, L. On-line learning and stochastic approximations. In *On-Line Learning in Neural Networks*, pp. 9–42. Cambridge University Press, 1 edition, January 1999. (page 1)

Bottou, L. Curiously fast convergence of some stochastic gradient descent algorithms. Unpublished open problem at the International Symposium on Statistical Learning and Data Sciences (SLDS), 2009. (page 9)

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, January 2018. (page 1)

Cha, J., Lee, J., and Yun, C. Tighter lower bounds for shuffling SGD: Random permutations and beyond. In *Proceedings of the International Conference on Machine Learning*, volume 202 of *PMLR*, pp. 3855–3912. JMLR, July 2023. (page 9)

Csiba, D. and Richtárik, P. Importance sampling for minibatches. *Journal of Machine Learning Research*, 19(27): 1–21, 2018. (pages 3, 37)

Dai, B., Xie, B., He, N., Liang, Y., Raj, A., Balcan, M.-F. F., and Song, L. Scalable kernel methods via doubly stochastic gradients. In *Advances in Neural Information Processing Systems*, volume 27, pp. 3041–3049. Curran Associates, Inc., 2014. (pages 1, 3, 9)

Domke, J. Provable gradient variance guarantees for black-box variational inference. In *Advances in Neural Information Processing Systems*, volume 32, pp. 329–338. Curran Associates, Inc., 2019. (pages 2, 37)

Domke, J. Provable smoothness guarantees for black-box variational inference. In *Proceedings of the International Conference on Machine Learning*, volume 119 of *PMLR*, pp. 2587–2596. JMLR, July 2020. (page 38)

Domke, J., Gower, R., and Garrigos, G. Provable convergence guarantees for black-box variational inference. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66289–66327. Curran Associates, Inc., 2023. (page 16)

Duchi, J. C., Bartlett, P. L., and Wainwright, M. J. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, January 2012. (page 35)

Garrigos, G. and Gower, R. M. Handbook of convergence theorems for (stochastic) gradient methods. Preprint arXiv:2301.11235, arXiv, February 2023. (pages 4, 19)

Gorbunov, E., Hanzely, F., and Richtarik, P. A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 108 of *PMLR*, pp. 680–690. JMLR, June 2020. (page 37)

Gower, R., Sebbouh, O., and Loizou, N. SGD for structured nonconvex functions: Learning rates, minibatching and interpolation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 130 of *PMLR*, pp. 1315–1323. JMLR, March 2021a. (pages 3, 4, 6, 15, 16, 21, 35)

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. SGD: General analysis and improved rates. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *PMLR*, pp. 5200–5209. JMLR, June 2019. (pages 2, 3, 4, 6, 16, 21, 37)

Gower, R. M., Schmidt, M., Bach, F., and Richtarik, P. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, November 2020. (page 9)

Gower, R. M., Richtárik, P., and Bach, F. Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching. *Mathematical Programming*, 188(1): 135–192, July 2021b. (pages 2, 3, 4)

Guminov, S., Gasnikov, A., and Kuruzov, I. Accelerated methods for weakly-quasi-convex optimization problems. *Computational Management Science*, 20(1): 36, December 2023. (pages 4, 16)

Gürbüzbalaban, M., Ozdaglar, A., and Parrilo, P. A. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186(1):49–84, March 2021. (page 9)

Haochen, J. and Sra, S. Random shuffling beats SGD after finite epochs. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *PMLR*, pp. 2624–2633. JMLR, May 2019. (page 9)

Hinder, O., Sidford, A., and Sohoni, N. Near-optimal methods for minimizing star-convex functions and beyond. In *Proceedings of Conference on Learning Theory*, volume 125 of *PMLR*, pp. 1894–1938. JMLR, July 2020. (page 16)

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. (page 1)

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, volume 26, pp. 315–323. Curran Associates, Inc., 2013. (page 3)

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pp. 795–811, Cham, 2016. Springer International Publishing. (pages 15, 38)

Khaled, A. and Richtárik, P. Better theory for SGD in the nonconvex world. *Transactions of Machine Learning Research*, 2023. (pages 4, 16)

Kim, K., Oh, J., Wu, K., Ma, Y., and Gardner, J. R. On the convergence of black-box variational inference. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44615–44657, New Orleans, LA, USA, December 2023. Curran Associates Inc. (pages 2, 38)

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations*, Banff, AB, Canada, April 2014. (pages 1, 37)

Kingma, D. P., Salimans, T., and Welling, M. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2575–2583. Curran Associates, Inc., 2015. (pages 2, 3)

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14): 1–45, 2017. (pages 1, 3, 37)

Kulunchakov, A. and Mairal, J. Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise. *Journal of Machine Learning Research*, 21(155):1–52, 2020. (pages 1, 9)

Liu, T., Li, Y., Wei, S., Zhou, E., and Zhao, T. Noisy gradient descent converges to flat minima for nonconvex matrix factorization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 130 of *PMLR*, pp. 1891–1899. JMLR, March 2021. (pages 1, 3, 9, 35)

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *Proceedings of the International Conference on Machine Learning*, volume 80 of *PMLR*, pp. 3325–3334. JMLR, July 2018. (pages 6, 35)

Mishchenko, K., Khaled, A., and Richtarik, P. Random reshuffling: Simple analysis with vast improvements. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17309–17320. Curran Associates, Inc., 2020. (pages 7, 8, 9, 30, 33)

Mohamed, S., Rosca, M., Figurnov, M., and Mnih, A. Monte Carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020. (pages 9, 37)

Moulines, E. and Bach, F. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, pp. 451–459. Curran Associates, Inc., 2011. (pages 2, 3, 9, 16)

Needell, D. and Ward, R. Batched stochastic gradient descent with weighted sampling. In *Approximation Theory XV: San Antonio 2016*, Springer Proceedings in Mathematics & Statistics, pp. 279–306, Cham, 2017. Springer International Publishing. (page 37)

Needell, D., Srebro, N., and Ward, R. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Mathematical Programming*, 155(1): 549–573, January 2016. (page 37)

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4): 1574–1609, January 2009. (pages 1, 9)

Nesterov, Y. and Polyak, B. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, August 2006. (page 16)

Nesterov, Yu. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, May 2005. (page 35)

Nguyen, L., Nguyen, P. H., van Dijk, M., Richtarik, P., Scheinberg, K., and Takac, M. SGD and Hogwild! Convergence without the bounded gradients assumption. In *Proceedings of the International Conference on Machine Learning*, volume 80 of *PMLR*, pp. 3750–3758. JMLR, July 2018. (pages 1, 3, 9, 15)

Nguyen, L. M., Tran-Dinh, Q., Phan, D. T., Nguyen, P. H., and Van Dijk, M. A unified convergence analysis for shuffling-type gradient methods. *The Journal of Machine Learning Research*, 22(1):207:9397–207:9440, January 2021. (page 9)

Orvieto, A., Raj, A., Kersting, H., and Bach, F. Explicit regularization in overparametrized models via noise injection. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 206 of *PMLR*, pp. 7265–7287. JMLR, April 2023. (pages 1, 3, 9, 35)

Polyak, B. T. and d Aleksandr Borisovich, T. Optimal order of accuracy of search algorithms in stochastic optimization. *Problemy Peredachi Informatsii*, 26(2):45–53, 1990. (page 35)

Ranganath, R., Gerrish, S., and Blei, D. Black box variational inference. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 33 of *PMLR*, pp. 814–822. JMLR, April 2014. (page 1)

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the International Conference on Machine Learning*, volume 32 of *PMLR*, pp. 1278–1286. JMLR, June 2014. (pages 1, 37)

Richtárik, P. and Takáč, M. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1-2):433–484, March 2016. (page 3)

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3): 400–407, September 1951. (page 1)

Safran, I. and Shamir, O. How good is SGD with random shuffling? In *Proceedings of Conference on Learning Theory*, volume 125 of *PMLR*, pp. 3250–3284. JMLR, July 2020. (page 9)

Safran, I. and Shamir, O. Random shuffling beats SGD only after many epochs on ill-conditioned problems. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15151–15161. Curran Associates, Inc., 2021. (page 9)

Schmidt, M. and Roux, N. L. Fast convergence of stochastic gradient descent under a strong growth condition. arXiv Preprint arXiv:1308.6370, arXiv, August 2013. (page 16)

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Stochastic convex optimization. In *Proceedings of the Conference on Computational Learning Theory*, June 2009. (page 9)

Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical Programming*, 127(1):3–30, March 2011. (page 1)

Shi, W., Gu, B., Li, X., Deng, C., and Huang, H. Triply stochastic gradient method for large-scale nonlinear similar unlabeled classification. *Machine Learning*, 110(8): 2005–2033, August 2021. (pages 1, 9)

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning*, volume 37 of *PMLR*, pp. 2256–2265. JMLR, June 2015. (page 1)

Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *Proceedings of the IEEE Global Conference on Signal and Information Processing*, pp. 245–248, Austin, TX, USA, December 2013. IEEE. (page 1)

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11918–11930. Curran Associates, Inc., 2019. (page 1)

Stich, S. U. Unified optimal analysis of the (stochastic) gradient method. Preprint arXiv:1907.04232, arXiv, December 2019. (page 4)

Titsias, M. and Lázaro-Gredilla, M. Doubly stochastic variational Bayes for non-conjugate inference. In *Proceedings of the International Conference on Machine Learning*, volume 32 of *PMLR*, pp. 1971–1979. JMLR, June 2014. (pages 1, 3, 37)

Vapnik, V. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, volume 4, pp. 831–838. Morgan-Kaufmann, 1991. (page 1)

Vaswani, S., Bach, F., and Schmidt, M. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 89 of *PMLR*, pp. 1195–1204. JMLR, April 2019. (pages 6, 16, 35)

Wright, S. J. and Recht, B. *Optimization for Data Analysis*. Cambridge University Press, New York, 2021. (page 16)

Xie, B., Liang, Y., and Song, L. Scale up nonlinear component analysis with doubly stochastic gradients. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2341–2349. Curran Associates, Inc., 2015. (page 9)

Xu, M., Quiroz, M., Kohn, R., and Sisson, S. A. Variance reduction properties of the reparameterization trick. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 89 of *PMLR*, pp. 2711–2720. JMLR, April 2019. (pages 9, 37)

Zheng, S. and Kwok, J. T.-Y. Lightweight stochastic optimization for minimizing finite sums with infinite data. In *Proceedings of the International Conference on Machine Learning*, volume 80 of *PMLR*, pp. 5932–5940. JMLR, July 2018. (pages 1, 9)

TABLE OF CONTENTS

1 Introduction **1**

2 Preliminaries **2**
2.1 Stochastic Gradient Descent on Finite-Sums . . . 2
2.2 Doubly Stochastic Gradients . . . 3
2.3 Technical Assumptions on Gradient Estimators . . . 3
2.4 Convergence Guarantees for SGD . . . 4

3 Main Results **5**
3.1 Doubly Stochastic Gradients . . . 5
3.1.1 General Variance Bound . . . 5
3.1.2 Gradient Variance Conditions for SGD . . . 6
3.2 Random Reshuffling of Stochastic Gradients . . . 7
3.2.1 Algorithm . . . 7
3.2.2 Proof Sketch . . . 7
3.2.3 Complexity Analysis . . . 8

4 Simulation **8**

5 Discussions **9**
5.1 Applications . . . 9
5.2 Related Works . . . 9
5.3 Conclusions . . . 9

A Gradient Variance Conditions **15**
A.1 Definitions . . . 15
A.2 Additional Gradient Variance Conditions . . . 15
A.3 Establishing the ER Condition . . . 16
A.4 Proofs of Implications in Fig. 2 . . . 16

B Proofs **18**
B.1 Auxiliary Lemmas (Lemmas 2 to 6) . . . 18
B.2 Convergence of SGD (Lemmas 1, 7 and 8) . . . 21
B.3 General Variance Bound (Theorem 1) . . . 23
B.4 Doubly Stochastic Gradients . . . 25
B.4.1 Expected Residual Condition (Theorem 2) . . . 25
B.4.2 Bounded Variance Condition (Theorem 3) . . . 27
B.4.3 Complexity Analysis (Corollary 2) . . . 27
B.5 Random Reshuffling of Stochastic Gradients . . . 29
B.5.1 Gradient Variance Conditions (Lemma 11, Lemma 12) . . . 29
B.5.2 Convergence Analysis (Theorem 5) . . . 31
B.5.3 Complexity Analysis (Theorem 4) . . . 34

C Applications **35**
C.1 ERM with Randomized Smoothing . . . 35
C.1.1 Description . . . 35
C.1.2 Preliminaries . . . 35
C.1.3 Theoretical Analysis . . . 36
C.2 Reparameterization Gradient . . . 37
C.2.1 Description . . . 37
C.2.2 Preliminaries . . . 37
C.2.3 Theoretical Analysis . . . 38

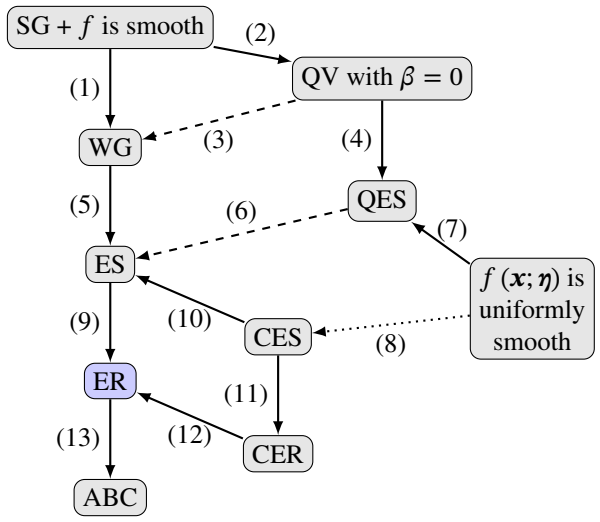


Figure 2. **Implications between general gradient variance conditions for some unbiased estimator $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x};\boldsymbol{\eta})$ of $\nabla f(\boldsymbol{x}) = \mathbb{E}\boldsymbol{g}(\boldsymbol{x})$.** The dashed arrows (- - ►) hold if f is further assumed to be QFG; the dotted arrow (·····►) holds if the integrand $f(\boldsymbol{x};\boldsymbol{\eta})$ is uniformly convex such that it is convex with respect to $\boldsymbol{x}$ for any fixed $\boldsymbol{\eta}$. (1), (5), (9), (13) are established by Gower et al. (2021a, Theorem 3.4); (2) is proven in Proposition 3; (3) is proven in Proposition 7; (4) is proven in Proposition 6; (7) is proven in Proposition 4; (8) is proven in Proposition 5; (6) is proven by Nguyen et al. (2018, Lemma 2) but we restate the proof in Proposition 8; (11) is proven in Proposition 2; (10), (12) hold trivially if $\boldsymbol{x}_* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$ are all stationary points.

A. Gradient Variance Conditions

In this section, we will discuss some additional aspects of the ER and ES conditions introduced in Section 2.3. We will also look into alternative gradient variance conditions that have been proposed in the literature and their relationship with the ER condition.

A.1. Definitions

For this section, we will use the following additional definitions:

Definition 3 (Quadratic Functional Growth; QFG)**.** We say $f : \mathcal{X} \to \mathbb{R}$ satisfies μ-quadratic functional growth if there exists some $\mu > 0$ such that

$$\frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2 \leq f(\boldsymbol{x}) - f(\boldsymbol{x}_*)$$

holds for all $\boldsymbol{x} \in \mathcal{X}$, where $\boldsymbol{x}_* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$.

This condition implies that f grows at least as fast as some quadratic and is weaker than the Polyak-Łojasiewicz. However, for any convex function f that satisfies this condition also means that f is μ-strongly convex (Karimi et al., 2016).

Definition 4 (Uniform Smoothness)**.** For the unbiased estimator $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x};\boldsymbol{\eta})$ of $\nabla F(\boldsymbol{x}) = \mathbb{E}\nabla f(\boldsymbol{x};\boldsymbol{\eta}) = \mathbb{E}\nabla f(\boldsymbol{x};\boldsymbol{\eta})$, we say the integrand $\nabla f_i(\boldsymbol{x};\boldsymbol{\eta})$ satisfies uniform L-smoothness if there exist some $L < \infty$ such that, for any fixed $\boldsymbol{\eta}$,

$$\left\|\nabla f(\boldsymbol{x};\boldsymbol{\eta}) - \nabla f(\boldsymbol{x}';\boldsymbol{\eta})\right\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{x}'\|_2$$

holds for all $(\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{X}^2$ simultaneously.

As discussed in Sections 1 and 5.2, this condition is rather strong: it does not hold for multiplicative noise unless the support is bounded.

Definition 5 (Uniform Convexity)**.** For the unbiased estimator $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x};\boldsymbol{\eta})$ of $\nabla F(\boldsymbol{x}) = \mathbb{E}\nabla f(\boldsymbol{x};\boldsymbol{\eta})$, we say the integrand $f(\boldsymbol{x};\boldsymbol{\eta})$ is uniformly convex if it is convex for any $\boldsymbol{\eta}$ such that, for any fixed $\boldsymbol{\eta}$,

$$f(\boldsymbol{x};\boldsymbol{\eta}) - f(\boldsymbol{x}';\boldsymbol{\eta}) \leq \langle \nabla f(\boldsymbol{x};\boldsymbol{\eta}), \boldsymbol{x} - \boldsymbol{x}' \rangle$$

holds for all $(\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{X}^2$ simultaneously.

A.2. Additional Gradient Variance Conditions

For some estimator $\boldsymbol{g}$ of ∇f, the following conditions have been considered in the literature:

- Strong growth condition (SG):
$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x})\|_2^2 \leq \rho\|\nabla f(\boldsymbol{x})\|_2^2$$
- Weak growth condition (WG):
$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x})\|_2^2 \leq \rho(f(\boldsymbol{x}) - f(\boldsymbol{x}_*))$$
- Quadratic variance (QV):
$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x})\|_2^2 \leq \alpha\|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2 + \beta$$
- Convex expected smoothness (CES):
$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y})\|_2^2 \leq 2\mathcal{L}\mathrm{D}_f(\boldsymbol{x}, \boldsymbol{y})$$
- Convex expected residual (CER):
$$\mathrm{tr}\mathbb{V}[\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y})] \leq 2\mathcal{L}\mathrm{D}_f(\boldsymbol{x}, \boldsymbol{y})$$
- Quadratic expected smoothness (QES):
$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y})\|_2^2 \leq \mathcal{L}^2\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$
- ABC:
$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x})\|_2^2 \leq A(f(\boldsymbol{x}) - f(\boldsymbol{x}_*)) + B\|\nabla f(\boldsymbol{x})\|_2^2 + C$$

Here, $\boldsymbol{x}_* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$ is any stationary point of f and the stated conditions should hold for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^2$.

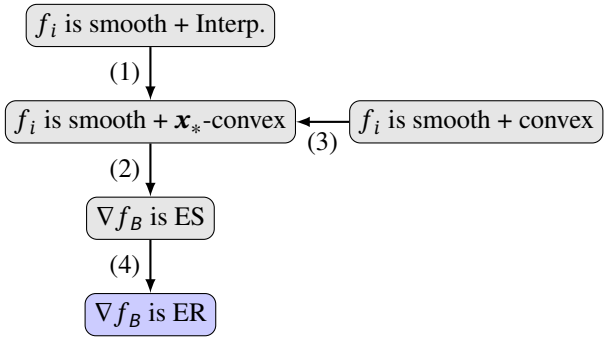


Figure 3. **Implications of assumptions on the components $f_1, \ldots, f_n$ to the minibatch subsampling gradient estimator ∇f_B of $F = \frac{1}{n}(f_1 + \ldots + f_n)$.** (1), (4) are established by Gower et al. (2021a, Theorem 3.4), while (3) trivially follows from the fact that $\boldsymbol{x}_*$-convexity is strictly weaker than (global) convexity, and (2) was established by Gower et al. (2019, Proposition 3.10).

SG was used by Schmidt & Roux (2013) to establish the linear convergence of SGD for strongly convex objectives, and $\mathcal{O}(1/T)$ convergence for convex objectives; **WG** was proposed by Vaswani et al. (2019) to establish similar guarantees to SG under a verifiable condition; **QV** was used to establish the non-asymptotic convergence on strongly convex functions by Wright & Recht (2021), while convergence on general convex functions was established by Domke et al. (2023), including stochastic proximal gradient descent; **QES** was used by (Moulines & Bach, 2011) to establish one of the earliest general non-asymptotic convergence results for SGD on strongly convex objectives; **ABC** was used by Khaled & Richtárik (2023) to establish convergence of SGD for non-convex smooth functions. (See also Khaled & Richtárik (2023) for a comprehensive overview of these conditions.) The relationship of these conditions with the ER condition are summarized in Fig. 2.

As demonstrated in Fig. 2 and discussed by Khaled & Richtárik (2023), the ABC condition is the weakest of all. However, the convergence guarantees for problems that exclusively satisfy the ABC condition are weaker than others. (For instance, the number of iterations T has to be fixed *a priori*.) On the other hand, the ER condition retrieves most of the strongest known guarantees for SGD; some of which were listed in Section 2.4.

A.3. Establishing the ER Condition

For subsampling estimators, it is possible to establish some of the gradient variance conditions through general assumptions on the components. See some examples in Fig. 3. Here, we use the following definitions:

Definition 6. For the finite sum objective $F = \frac{1}{n}(f_1 + \ldots + f_n)$, we say interpolation holds if, for all $i = 1, \ldots, n$,

$$f_i(\boldsymbol{x}_*) \leq f_i(\boldsymbol{x}),$$

holds for all $\boldsymbol{x} \in \mathcal{X}$, where $\boldsymbol{x}_* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$.

Definition 7. For the finite sum objective $F = \frac{1}{n}(f_1 + \ldots + f_n)$, we say the components are $\boldsymbol{x}_*$-convex if, for all $i = 1, \ldots, n$,

$$f_i(\boldsymbol{x}_*) - f_i(\boldsymbol{x}) \leq \langle \nabla f_i(\boldsymbol{x}_*), \boldsymbol{x}_* - \boldsymbol{x} \rangle$$

holds for all $\boldsymbol{x} \in \mathcal{X}$, where $\boldsymbol{x}_* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$.

This assumption is a weaker version of convexity; convexity needs to hold with respect to $\boldsymbol{x}_*$ only. It is closely related to star (Nesterov & Polyak, 2006) and quasar convexity (Hinder et al., 2020; Guminov et al., 2023).

A.4. Proofs of Implications in Fig. 2

We prove new implication results between some of the gradient variance conditions discussed in Appendix A.2. In particular, the relationship between the QES and QV against other conditions has not been considered before.

Proposition 2. *Let $\boldsymbol{g}$ be an unbiased estimator of ∇f. Then,*

$$\boldsymbol{g} \text{ is CES} \Longrightarrow \boldsymbol{g} \text{ is CER}$$

Proof. The result immediately follows from the fact that

$$\operatorname{tr}\mathbb{V}\left[\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x}')\right] \leq \mathbb{E}\left\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x}')\right\|_2^2$$

holds for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$. □

Proposition 3. *Let $\boldsymbol{g}$ be an unbiased estimator of ∇f. Then,*

$$\boldsymbol{g} \text{ is SG} + f \text{ is } L\text{-smooth} \Longrightarrow \boldsymbol{g} \text{ is QV with } \beta = 0$$

Proof. Notice that, by definition, $\nabla f(\boldsymbol{x}_*) = 0$. Then,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x})\|_2^2 &\leq \rho\|\nabla f(\boldsymbol{x})\|_2^2 \\ &= \rho\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}_*)\|_2^2,\end{aligned}$$

applying L-smoothness of f,

$$\leq L^2\rho\|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2.$$

□

Proposition 4. *Let $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x}; \boldsymbol{\eta})$ be an unbiased estimator of $\nabla f(\boldsymbol{x}) = \mathbb{E}\nabla f(\boldsymbol{x}; \boldsymbol{\eta})$. Then,*

Integrand is uniformly L-smooth $\Longrightarrow$ *QES*

Proof. The result immediately follows from the fact that the integrand $f(\boldsymbol{x}; \boldsymbol{\eta})$ is L-smooth with respect to $\boldsymbol{x}$ uniformly over $\boldsymbol{\eta}$ as

$$\mathbb{E}\left\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x}')\right\|_2^2 = \mathbb{E}\left\|\nabla f(\boldsymbol{x}; \boldsymbol{\eta}) - \nabla f(\boldsymbol{x}'; \boldsymbol{\eta})\right\|_2^2 \leq L^2 \|\boldsymbol{x} - \boldsymbol{x}'\|_2^2.$$

□

Proposition 5. *Let $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x}; \boldsymbol{\eta})$ be an unbiased estimator of $\nabla f(\boldsymbol{x}) = \mathbb{E}\nabla f(\boldsymbol{x}; \boldsymbol{\eta})$. Then,*

Integrand is uniformly L-smooth + *f is uniformly convex* $\Longrightarrow$ *ES*

Proof. Since the integrand $f(\boldsymbol{x}; \boldsymbol{\eta})$ is both uniformly smooth and convex with respect to $\boldsymbol{x}$ for a any fixed $\boldsymbol{\eta}$, we have

$$\left\|\nabla f(\boldsymbol{x}; \boldsymbol{\eta}) - \nabla f(\boldsymbol{x}'; \boldsymbol{\eta})\right\|_2 \leq 2L\left(f(\boldsymbol{x}; \boldsymbol{\eta}) - f(\boldsymbol{x}'; \boldsymbol{\eta}) - \left\langle \nabla f(\boldsymbol{x}'; \boldsymbol{\eta}), \boldsymbol{x} - \boldsymbol{x}' \right\rangle\right).$$

Then,

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x}_*)\|_2^2 \\
&= \mathbb{E}\|\nabla f(\boldsymbol{x}; \boldsymbol{\eta}) - \nabla f(\boldsymbol{x}_*; \boldsymbol{\eta})\|_2^2 \\
&\leq 2L\,\mathbb{E}\left(f(\boldsymbol{x}; \boldsymbol{\eta}) - f(\boldsymbol{x}_*; \boldsymbol{\eta}) - \left\langle \nabla f(\boldsymbol{x}_*; \boldsymbol{\eta}), \boldsymbol{x} - \boldsymbol{x}' \right\rangle\right) \\
&= 2L\left(f(\boldsymbol{x}) - f(\boldsymbol{x}_*) - \left\langle \nabla f(\boldsymbol{x}_*), \boldsymbol{x} - \boldsymbol{x}' \right\rangle\right) \\
&= 2L\left(f(\boldsymbol{x}) - f(\boldsymbol{x}_*)\right)
\end{aligned}$$

holds for all $\boldsymbol{x} \in \mathcal{X}$. □

Proposition 6. *Let $\boldsymbol{g}$ be an unbiased estimator of ∇F. Then,*

$\boldsymbol{g}$ is QV with $\beta = 0$ $\Longrightarrow$ *QES*

Proof. From the classic inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x}_*)\|_2^2 \leq 2\,\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x})\|_2^2 + 2\,\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}_*)\|_2^2.$$

Now, since QV holds with $\beta = 0$, we have $\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}_*)\|_2^2 = 0$. Therefore,

$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x}_*)\|_2^2 \leq 2\,\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x})\|_2^2 \leq 2\,\alpha\|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2,$$

where we have applied QV at the last inequality. □

Proposition 7. *Let $\boldsymbol{g}$ be an unbiased estimator of ∇f. Then,*

$\boldsymbol{g}$ is QV with $\beta = 0$ + *f is μ-QFG* $\Longrightarrow$ *WG*

Proof. The result immediately follows from QV as

$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x})\|_2^2 \leq \alpha\|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2,$$

applying μ-quadratic functional growth,

$$\leq \frac{2\alpha}{\mu}\left(f(\boldsymbol{x}) - f(\boldsymbol{x}_*)\right).$$

□

Proposition 8. *Let $\boldsymbol{g}$ be an unbiased estimator of ∇f. Then,*

$\boldsymbol{g}$ is QES + *f is μ-QFG* $\Longrightarrow$ *ES*

Proof. From QV, we have

$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x}_*)\|_2^2 \leq \mathcal{L}^2\|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2$$

and μ-quadratic functional growth yields

$$\leq \frac{2\mathcal{L}^2}{\mu}\left(f(\boldsymbol{x}) - f(\boldsymbol{x}_*)\right).$$

□

The strategy applying QFG when proving Propositions 7 and 8 establishes the stronger variant of the ER condition: Assumption 6 (B). However, the price for this is that one has to pay for an excess $\kappa = \mathcal{L}/\mu$ factor, and this strategy works only works for quadratically growing objectives.

B. Proofs

B.1. Auxiliary Lemmas (Lemmas 2 to 6)

Lemma 2. *Consider a finite population of n vector-variate random variables $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$. Then, the variance of the average of b samples chosen without replacement is*

$$\mathrm{tr}\mathbb{V}\left[\frac{1}{b}\sum_{i=1}^{b}\boldsymbol{x}_{B_i}\right] = \frac{n-b}{b\,(n-1)}\sigma^2,$$

where $B = \{B_1, \dots, B_b\}$ is the collection of random indices of the samples and σ^2 is the variance of independently choosing a single sample.

Proof. From the variance of the sum of random variables, we have

$$\mathrm{tr}\mathbb{V}\left[\sum_{i=1}^{b}\boldsymbol{x}_{B_i}\right] = \sum_{i=1}^{b}\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_{B_i}\right] + \sum_{i=1}^{b}\sum_{i\neq j}^{b}\mathrm{Cov}\left(\boldsymbol{x}_{B_i}, \boldsymbol{x}_{B_j}\right),$$

and noticing that the covariance is independent of the index in the batch,

$$= b\,\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_{B_i}\right] + b(b-1)C, \tag{12}$$

where $C = \mathrm{Cov}\left(\boldsymbol{x}_{B_i}, \boldsymbol{x}_{B_j}\right)$. Using the fact that the variance is 0 for $b = n$, we can solve for C such that

$$C = -\frac{1}{n-1}\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_{B_i}\right],$$

which is negative, and a negative correlation is always great. Plugging this expression to Eq. (12), we have

$$\begin{aligned}\mathrm{tr}\mathbb{V}\left[\sum_{i=1}^{b}\boldsymbol{x}_{B_i}\right] &= b\,\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_{B_i}\right] - b(b-1)\frac{1}{n-1}\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_{B_i}\right],\\ &= b\left(1 - \frac{b-1}{n-1}\right)\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_{B_i}\right]\\ &= b\left(\frac{n-b}{n-1}\right)\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_{B_i}\right].\end{aligned}$$

Dividing both sides by b^2 yields the result. □

Lemma 3. *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ be vector-variate random variables. Then, the variance of the sum is upper-bounded as*

$$\mathrm{tr}\mathbb{V}\left[\sum_{i=1}^{n}\boldsymbol{x}_i\right] \leq \left(\sum_{i=1}^{n}\sqrt{\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]}\right)^2 \tag{13}$$

$$\leq n\sum_{i=1}^{n}\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]. \tag{14}$$

The equality in Eq. (13) holds if and only if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are constant multiples such that there exists some $\alpha_{ij} \geq 0$ such that

$$\boldsymbol{x}_i = \alpha_{ij}\boldsymbol{x}_j$$

for all i, j.

Proof. The variance of a sum is

$$\mathrm{tr}\mathbb{V}\left[\sum_{i=1}^{n}\boldsymbol{x}_i\right] = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{tr}\,\mathrm{Cov}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right).$$

From the Cauchy-Schwarz inequality for expectations,

$$\begin{aligned}\mathrm{tr}\,\mathrm{Cov}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) &= \mathbb{E}(\boldsymbol{x}_i - \mathbb{E}\boldsymbol{x}_i)^\top(\boldsymbol{x}_j - \mathbb{E}\boldsymbol{x}_j)\\ &\leq \mathbb{E}\|\boldsymbol{x}_i - \mathbb{E}\boldsymbol{x}_i\|_2\mathbb{E}\|\boldsymbol{x}_j - \mathbb{E}\boldsymbol{x}_j\|_2\\ &= \sqrt{\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]}\sqrt{\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_j\right]}.\end{aligned}$$

This implies

$$\begin{aligned}\mathrm{tr}\mathbb{V}\left[\sum_{i=1}^{n}\boldsymbol{x}_i\right] &= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{tr}\,\mathrm{Cov}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\\ &\leq \sum_{i=1}^{n}\sum_{j=1}^{n}\sqrt{\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]}\sqrt{\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_j\right]}\\ &= \left(\sum_{i=1}^{n}\sqrt{\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]}\right)^2. \end{aligned}\tag{15}$$

The equality statement comes from the property of the Cauchy-Schwarz inequality. Lastly, Eq. (14) follows from additionally applying Jensen's inequality as

$$\begin{aligned}\left(\sum_{i=1}^{n}\sqrt{\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]}\right)^2 &= n^2\left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]}\right)^2\\ &\leq n^2\frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]}\right)^2\\ &= n\sum_{i=1}^{n}\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_i\right].\end{aligned}$$

An equivalent proof strategy is to expand the quadratic in Eq. (15) and apply the arithmetic mean-geometric mean inequality to the cross terms. □

Lemma 4 (Lemma A.2; Garrigos & Gower, 2023)**.** *For a recurrence relation given as*

$$r_T \leq (1-\gamma\mu)^T r_0 + B\gamma,$$

for some constant $0 < \gamma < 1/C$,

$$r_T \leq \epsilon$$

can be guaranteed by setting

$$\gamma = \min\left(\frac{\epsilon}{2B}, \frac{1}{C}\right) \text{ and}$$
$$T \geq \frac{1}{\mu}\max\left(2B\frac{1}{\epsilon}, C\right)\log\left(2\frac{r_0}{\epsilon}\right),$$

where $\mu, B > 0$ *and* $0 < C < \mu$ *are some finite constants.*

Proof. First, notice that the recurrence

$$r_T \leq \underbrace{(1-\gamma\mu)^T r_0}_{\text{bias}} + \underbrace{B\gamma}_{\text{variance}},$$

is a sum of monotonically increasing (variance) and decreasing (bias) terms with respect to γ. Therefore, the bound is minimized when both terms are equal. This implies that $r_t \leq \epsilon$ can be achieved by solving for

$$(1-\gamma\mu)^T r_0 \leq \frac{\epsilon}{2} \quad \text{and} \quad B\gamma \leq \frac{\epsilon}{2}$$

First, for the variance term,

$$B\gamma \leq \frac{\epsilon}{2} \quad \Leftrightarrow \quad \gamma \leq \frac{\epsilon}{2B}.$$

For the bias term, as long as $\gamma < \frac{1}{\mu}$,

$$\begin{aligned}
&(1-\gamma\mu)^T r_0 \leq \frac{\epsilon}{2} \\
\Leftrightarrow \quad &T\log(1-\gamma\mu) \leq \log\frac{\epsilon}{2r_0} \\
\Leftrightarrow \quad &T \leq \frac{\log\frac{\epsilon}{2r_0}}{\log(1-\gamma\mu)} \\
\Leftrightarrow \quad &T \geq \frac{\log\frac{2r_0}{\epsilon}}{\log(1/(1-\gamma\mu))}
\end{aligned}$$

Furthermore, using the bound $\log 1/x \geq 1-x$ for $0 < x < 1$, we can achieve the guarantee with

$$T \geq \frac{1}{\gamma\mu}\log\left(\frac{2r_0}{\epsilon}\right).$$

Therefore, $1/\gamma$ determines the iteration complexity. Plugging in the minimum over the constraints on γ yields the iteration complexity. □

Lemma 5. *For a recurrence relation given as*

$$r_T \leq (1-\gamma\mu)^T r_0 + A\gamma^2 + B\gamma,$$

for some constant $0 < \gamma < 1/C$,

$$r_T \leq \epsilon$$

can be guaranteed by setting

$$\gamma = \min\left(\frac{-B+\sqrt{B^2+2A\epsilon}}{2A}, \frac{1}{C}\right) \text{ and}$$
$$T \geq \frac{1}{\mu}\max\left(2B\frac{1}{\epsilon} + \sqrt{2A}\frac{1}{\sqrt{\epsilon}}, C\right)\log\left(2\frac{r_0}{\epsilon}\right),$$

where $\mu, A, B > 0$ *and* $0 < C < \mu$ *are some finite constants.*

Proof. This theorem is a generalization of Lemma A.2 by Garrigos & Gower (2023). First, notice that the recurrence

$$r_T \leq \underbrace{(1-\gamma\mu)^T r_0}_{\text{bias}} + \underbrace{A\gamma^2 + B\gamma}_{\text{variance}},$$

is a sum of monotonically increasing (variance) and decreasing (bias) terms with respect to γ. Therefore, the bound is minimized when both terms are equal. This implies that $r_t \leq \epsilon$ can be achieved by solving for

$$(1-\gamma\mu)^T r_0 \leq \frac{\epsilon}{2} \quad \text{and} \quad A\gamma^2 + B\gamma \leq \frac{\epsilon}{2}$$

First, for the variance term,

$$\begin{aligned}
&A\gamma^2 + B\gamma \leq \frac{\epsilon}{2} \\
\Leftrightarrow \quad &A\gamma^2 + B\gamma - \frac{\epsilon}{2} \leq 0
\end{aligned}$$

The solution to this equation is given by the positive solution of the quadratic equation as

$$0 < \gamma \leq \frac{-B+\sqrt{B^2+2A\epsilon}}{2A}.$$

For the bias term, as long as $\gamma < \frac{1}{\mu}$, the solution is identical to Lemma 4. Therefore,

$$T \geq \frac{1}{\gamma\mu}\log\left(\frac{2r_0}{\epsilon}\right) \tag{16}$$

can guarantee the bias term to be smaller than $\epsilon/2$, while $1/\gamma$ determines the iteration complexity. Plugging in the minimum over the constraints on γ,

$$\gamma = \min\left(\frac{-B+\sqrt{B^2+2A\epsilon}}{2A}, \frac{1}{C}\right) \tag{17}$$

yields the iteration complexity.

Now, since the quadratic formula is not very interpretable, let us simplify the expression for $1/\gamma$ using the bound

$$\frac{a}{2\sqrt{b^2+a}} \leq -b + \sqrt{b^2+a},$$

which holds for any $a, b > 0$ and is tight for $\epsilon \to 0$. With our constants, this reads

$$\frac{A\epsilon}{\sqrt{B^2+2A\epsilon}} \leq -B + \sqrt{B^2+2A\epsilon},$$

and therefore

$$\begin{aligned}\frac{2A}{-B+\sqrt{B^2+2A\epsilon}} &\leq \frac{2\sqrt{B^2+2A\epsilon}}{\epsilon} \\ &\leq \frac{2B+\sqrt{2A\epsilon}}{\epsilon} \\ &= 2B\frac{1}{\epsilon} + \sqrt{2A}\frac{1}{\sqrt{\epsilon}}.\end{aligned}$$

Therefore, for the stepsize choice of Eq. (17),

$$\frac{1}{\gamma} \leq \min\left(2B\frac{1}{\epsilon} + \sqrt{2A}\frac{1}{\sqrt{\epsilon}}, \frac{1}{C}\right).$$

Plugging this into Eq. (16) yields the statement. □

Lemma 6. *Let $F : \mathcal{X} \to \mathbb{R}$ be a finite sum of convex functions as $F = \frac{1}{n}(f_1 + ... + f_n)$, where $f_i : \mathcal{X} \to \mathbb{R}$. Then,*

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{D}_{f_i}(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{D}_F(\boldsymbol{x}, \boldsymbol{x}'),$$

for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$.

Proof. The result immediately follows from the definition of Bregman divergences as

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} \mathrm{D}_{f_i}(\boldsymbol{x}, \boldsymbol{x}') \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(f_i(\boldsymbol{x}) - f_i(\boldsymbol{x}') - \langle \nabla f_i(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}'\rangle\right) \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x})\right) - \left(\frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}')\right) \\ &\quad - \left\langle \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}'\right\rangle \\ &= F(\boldsymbol{x}) - F(\boldsymbol{x}') - \langle \nabla F(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}'\rangle \\ &= \mathrm{D}_F(\boldsymbol{x}, \boldsymbol{x}').\end{aligned}$$

□

B.2. Convergence of SGD (Lemmas 1, 7 and 8)

Lemma 7. *Let $F : \mathcal{X} \to \mathbb{R}$ be L-smooth function. Then, the expected squared norm of a gradient estimator $\boldsymbol{g}$ satisfying both $\mathrm{ER}(\mathcal{L})$ and $\mathrm{BV}(\sigma^2)$ is bounded as*

$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x})\|_2^2 \leq 4(\mathcal{L}+L)(F(\boldsymbol{x})-F(\boldsymbol{x}_*))+2\sigma^2,$$

for any $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{x}_* \in \arg\max_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$.

Proof. The proof is a minor modification of Lemma 2.4 by Gower et al. (2019) and Lemma 3.2 by Gower et al. (2021a).

By applying the bound $(a+b)^2 \leq 2a^2+2b^2$, we can "transfer" the variance on $\boldsymbol{x}$ to the variance of $\boldsymbol{x}_*$. That is,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x})\|_2^2 &= \mathbb{E}\|\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{x}_*)+\boldsymbol{g}(\boldsymbol{x}_*)\|_2^2 \\ &\leq 2\underbrace{\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{x}_*)\|_2^2}_{V_1}+2\underbrace{\mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}_*)\|_2^2}_{V_2}\end{aligned}$$

The key is to bound V_1. It is typical to do this using expected-smoothness-type assumptions such as the ER assumption. That is,

$$\begin{aligned}V_1 &= \mathbb{E}\|\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{x}_*)\|_2^2 \\ &= \mathrm{tr}\mathbb{V}[\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{x}_*)]+(\nabla F(\boldsymbol{x})-\nabla F(\boldsymbol{x}_*)),\end{aligned}$$

from the L-smoothness of F,

$$\leq \mathrm{tr}\mathbb{V}[\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{x}_*)]+2L(F(\boldsymbol{x})-F(\boldsymbol{x}_*)),$$

and the ER condition,

$$\begin{aligned}&\leq 2\mathcal{L}(F(\boldsymbol{x})-F(\boldsymbol{x}_*))+2L(F(\boldsymbol{x})-F(\boldsymbol{x}_*)) \\ &= 2(L+\mathcal{L})\big(F(\boldsymbol{x})-F(\boldsymbol{x}_*)\big).\end{aligned}$$

Finally, V_2 immediately follows from the BV condition as

$$V_2 = \mathbb{E}\|\boldsymbol{g}(\boldsymbol{x}_*)\|_2^2 \leq \sigma^2.$$

□

Lemma 8. *Let the objective function F satisfy Assumption 1 and the gradient estimator $\boldsymbol{g}$ be unbiased and satisfy both $\mathrm{ER}(\mathcal{L})$ and $\mathrm{BV}(\sigma^2)$. Then, the last iterate of SGD guarantees*

$$\mathbb{E}\left[\|\boldsymbol{x}_T-\boldsymbol{x}^*\|_2^2\right] \leq (1-\mu\gamma)^T\|\boldsymbol{x}_0-\boldsymbol{x}_*\|_2^2+\frac{2\sigma^2}{\mu}\gamma$$

where $\boldsymbol{x}_* = \arg\min_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$ is the global optimum.

Proof. Firstly, we have

$$\begin{aligned}&\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_*\|_2^2 \\ &= \|\Pi_{\mathcal{X}}(\boldsymbol{x}_t-\gamma\boldsymbol{g}(\boldsymbol{x}_t))-\Pi(\boldsymbol{x}_*)\|_2^2,\end{aligned}$$

and since the projection onto a convex set under a Euclidean metric is non-expansive,

$$\begin{aligned}&\leq \|\boldsymbol{x}_t-\gamma\boldsymbol{g}(\boldsymbol{x}_t)-\boldsymbol{x}_*\|_2^2 \\ &= \|\boldsymbol{x}_t-\boldsymbol{x}_*\|_2^2-2\gamma\langle\boldsymbol{g}(\boldsymbol{x}_t),\boldsymbol{x}_t-\boldsymbol{x}_*\rangle+\gamma^2\|\boldsymbol{g}(\boldsymbol{x}_t)\|_2^2.\end{aligned}$$

Denoting the σ-algebra formed by the randomness and the iterates up to the tth iteration as $\mathcal{F}_t$ such that $(\mathcal{F}_t)_{t\geq 1}$ forms a filtration, the conditional expectation is

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_*\|_2^2 \mid \mathcal{F}_t\right] \\ &= \|\boldsymbol{x}_t-\boldsymbol{x}_*\|_2^2-2\gamma\langle\mathbb{E}[\boldsymbol{g}(\boldsymbol{x}_t)\mid\mathcal{F}_t],\boldsymbol{x}_t-\boldsymbol{x}_*\rangle \\ &\quad+\gamma^2\mathbb{E}\left[\|\boldsymbol{g}(\boldsymbol{x}_t)\|_2^2\mid\mathcal{F}_t\right]. \\ &= \|\boldsymbol{x}_t-\boldsymbol{x}_*\|_2^2-2\gamma\langle\nabla F(\boldsymbol{x}_t),\boldsymbol{x}_t-\boldsymbol{x}_*\rangle \\ &\quad+\gamma^2\mathbb{E}\left[\|\boldsymbol{g}(\boldsymbol{x}_t)\|_2^2\mid\mathcal{F}_t\right],\end{aligned}$$

applying the μ-strong convexity of F,

$$\begin{aligned}&\leq \|\boldsymbol{x}_t-\boldsymbol{x}_*\|_2^2-2\gamma\left(F(\boldsymbol{x}_t)-F(\boldsymbol{x}_*)+\frac{\mu}{2}\|\boldsymbol{x}_t-\boldsymbol{x}_*\|_2^2\right) \\ &\quad+\gamma^2\mathbb{E}\left[\|\boldsymbol{g}(\boldsymbol{x}_t)\|_2^2\mid\mathcal{F}_t\right] \\ &= (1-\gamma\mu)\|\boldsymbol{x}_t-\boldsymbol{x}_*\|_2^2-2\gamma(F(\boldsymbol{x}_t)-F(\boldsymbol{x}_*)) \\ &\quad+\gamma^2\mathbb{E}\left[\|\boldsymbol{g}(\boldsymbol{x}_t)\|_2^2\mid\mathcal{F}_t\right]\end{aligned}$$

From Lemma 7, we have

$$\mathbb{E}\left[\|\boldsymbol{g}(\boldsymbol{x}_t)\|_2^2\mid\mathcal{F}_t\right] \leq \left(4(\mathcal{L}+L)(F(\boldsymbol{x}_t)-F(\boldsymbol{x}_*))+2\sigma^2\right).$$

Therefore,

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_*\|_2^2\mid\mathcal{F}_t\right] \\ &\leq (1-\gamma\mu)\|\boldsymbol{x}_t-\boldsymbol{x}_*\|_2^2-2\gamma(F(\boldsymbol{x}_t)-F(\boldsymbol{x}_*)) \\ &\quad+\gamma^2\left(4(\mathcal{L}+L)(F(\boldsymbol{x}_t)-F(\boldsymbol{x}_*))+2\sigma^2\right) \\ &= (1-\gamma\mu)\|\boldsymbol{x}_t-\boldsymbol{x}_*\|_2^2 \\ &\quad-2\gamma(1-2\gamma(\mathcal{L}+L))(F(\boldsymbol{x}_t)-F(\boldsymbol{x}_*))+2\gamma^2\sigma^2,\end{aligned}$$

and with a small-enough stepsize satisfying $\gamma < \frac{1}{2(\mathcal{L}+L)}$, we can guarantee a partial contraction as

$$\leq (1-\gamma\mu)\,\|\boldsymbol{x}_t - \boldsymbol{x}_*\|_2^2 + 2\gamma^2\sigma^2.$$

Note that the coefficient $1-\gamma\mu$ is guaranteed to be strictly smaller than 1 since $\mu \leq L$, which means that we indeed have a partial contraction.

Now, taking full expectation, we have

$$\mathbb{E}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_*\|_2^2 \leq (1-\gamma\mu)\,\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}_*\|_2^2 + 2\gamma^2\sigma^2.$$

Unrolling the recursion from 0 to $T-1$, we have

$$\begin{aligned}\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}_*\|_2^2 &\leq (1-\gamma\mu)^T\,\mathbb{E}\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2 \\ &\quad + 2\gamma^2\sigma^2\sum_{t=0}^{T-1}(1-\gamma\mu)^t. \\ &\leq (1-\gamma\mu)^T\,\mathbb{E}\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2 + \frac{2\sigma^2}{\mu}\gamma.\end{aligned}$$

where the last inequality follows from the asymptotic bound on geometric sums. □

Lemma 1. *Let the objective F satisfy Assumption 1 and the gradient estimator $\boldsymbol{g}$ satisfy $\mathrm{ER}(\mathcal{L})$ and $\mathrm{BV}(\sigma^2)$. Then, the last iterate of SGD is ϵ-close to the global optimum $\boldsymbol{x}_* = \arg\min_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$ such that $\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}_*\|_2^2 \leq \epsilon$ after a number of iterations at least*

$$T \geq 2\max\left(\frac{\sigma^2}{\mu^2}\frac{1}{\epsilon}, \frac{\mathcal{L}+L}{\mu}\right)\log\left(2\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2\frac{1}{\epsilon}\right)$$

and the fixed stepsize

$$\gamma = \min\left(\frac{\epsilon\mu}{2\sigma^2}, \frac{1}{2(\mathcal{L}+L)}\right).$$

Proof. We can apply Lemma 4 to the result of Lemma 8 with the constants

$$r_0 = \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2, \quad B = \frac{2\sigma^2}{\mu}, \text{ and } \; C = 2(\mathcal{L}+L).$$

Then, we can guarantee an ϵ-accurate solution with the stepsize

$$\gamma = \min\left(\frac{\epsilon\mu}{2\sigma^2}, \frac{1}{2(\mathcal{L}+L)}\right)$$

and a number of iterations of at least

$$\begin{aligned}T &\geq \frac{1}{\mu}\max\left(\frac{2\sigma^2}{\mu}, 2(\mathcal{L}+L)\right)\log\left(2\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2\frac{1}{\epsilon}\right) \\ &= 2\max\left(\frac{\sigma^2}{\mu^2}, \frac{\mathcal{L}+L}{\mu}\right)\log\left(2\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2\frac{1}{\epsilon}\right).\end{aligned}$$

□

B.3. General Variance Bound (Theorem 1)

Lemma 9. *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_b$ be a collection of vector-variate RVs dependent on some random variable B satisfying Assumption 3. Then, the expected variance of the sum of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_b$ conditioned on B is bounded as*

$$\mathbb{E}\left[\mathrm{tr}\mathbb{V}\left[\sum_{i=1}^b \boldsymbol{x}_i \mid B\right]\right] \leq \rho\mathbb{V}[S] + \rho(\mathbb{E}S)^2 + (1-\rho)\,\mathbb{E}[V],$$

where

$$S = \sum_{i=1}^b \sqrt{\mathrm{tr}\mathbb{V}[\boldsymbol{x}_i \mid B]} \quad \text{and} \quad V = \sum_{i=1}^b \mathrm{tr}\mathbb{V}[\boldsymbol{x}_i \mid B].$$

Equality holds when the equality in Assumption 3 holds.

Proof. From the formula for the variance of sums,

$$\begin{aligned}
&\mathrm{tr}\mathbb{V}\left[\sum_{i=1}^b \boldsymbol{x}_i \,\middle|\, B\right] \\
&= \sum_{i=1}^b \mathrm{tr}\,\mathbb{V}[\boldsymbol{x}_i \mid B] + \sum_{i=1}^b \sum_{j\neq i} \mathrm{tr}\,\mathrm{Cov}\left(\boldsymbol{x}_i, \boldsymbol{x}_j \mid B\right). \\
&\leq \sum_{i=1}^b \mathrm{tr}\,\mathbb{V}[\boldsymbol{x}_i \mid B] + \sum_{i=1}^b \sum_{j\neq i} \rho\sqrt{\mathrm{tr}\mathbb{V}[\boldsymbol{x}_i \mid B]}\sqrt{\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_j \mid B\right]} \\
&= (1-\rho)\sum_{i=1}^b \mathrm{tr}\,\mathbb{V}[\boldsymbol{x}_i \mid B] \\
&\quad + \rho\sum_{i=1}^b \sum_{j=1}^b \sqrt{\mathrm{tr}\mathbb{V}[\boldsymbol{x}_i \mid B]}\sqrt{\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}_j \mid B\right]} \\
&= (1-\rho)\sum_{i=1}^b \mathrm{tr}\,\mathbb{V}[\boldsymbol{x}_i \mid B] + \rho\left(\sum_{i=1}^b \sqrt{\mathrm{tr}\mathbb{V}[\boldsymbol{x}_i \mid B]}\right)^2 \\
&= (1-\rho)\,V + \rho\,S^2.
\end{aligned}$$

Then, it follows that

$$\begin{aligned}
\mathbb{E}\left[\mathrm{tr}\mathbb{V}\left[\sum_{i=1}^b \boldsymbol{x}_i \,\middle|\, B\right]\right] &\leq \rho\mathbb{E}\left[S^2\right] + (1-\rho)\,\mathbb{E}[V] \\
&= \rho\mathbb{V}[S] + \rho(\mathbb{E}S)^2 + (1-\rho)\,\mathbb{E}[V],
\end{aligned}$$

from the basic property of the variance:

$$\mathbb{V}[S] = \mathbb{E}\left[S^2\right] - (\mathbb{E}S)^2.$$

Since Assumption 3 is the only inequality we use, the equality in the statement holds whenever the equality in Assumption 3 holds. □

Theorem 1. *Let the component estimators $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ satisfy Assumption 3. Then, the variance of the doubly stochastic estimator $\boldsymbol{x}_B$ is bounded as*

$$\mathrm{tr}\mathbb{V}[\boldsymbol{x}_B] \leq V_{\mathrm{com}} + V_{\mathrm{cor}} + V_{\mathrm{sub}},$$

where

$$\begin{aligned}
V_{\mathrm{com}} &= \left(\frac{\rho}{b_{\mathrm{eff}}} + \frac{1-\rho}{b}\right)\left(\frac{1}{n}\sum_{i=1}^n \mathrm{tr}\mathbb{V}[\boldsymbol{x}_i]\right), \\
V_{\mathrm{cor}} &= \rho\left(1 - \frac{1}{b_{\mathrm{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^n \sqrt{\mathrm{tr}\mathbb{V}[\boldsymbol{x}_i]}\right)^2, \text{ and} \\
V_{\mathrm{sub}} &= \frac{1}{b_{\mathrm{eff}}}\frac{1}{n}\sum_{i=1}^n \|\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}\|_2^2.
\end{aligned}$$

Equality holds when the equality in Assumption 3 holds.

Proof. Starting from the law of total covariance, we have

$$\mathbb{V}[\boldsymbol{x}_B] = \underbrace{\mathbb{E}_{B\sim\pi}[\mathrm{tr}\mathbb{V}[\boldsymbol{x}_B \mid B]]}_{\text{Ensemble Variance}} + \underbrace{\mathrm{tr}\mathbb{V}_{B\sim\pi}[\mathbb{E}[\boldsymbol{x}_B \mid B]]}_{\text{Subsampling Variance}}. \quad (18)$$

Ensemble Variance Bounding the variance of each ensemble is key. From Lemma 9, we have

$$\begin{aligned}
\mathbb{E}[\mathrm{tr}\mathbb{V}[\boldsymbol{x}_B \mid B]] &= \mathbb{E}\left[\mathrm{tr}\mathbb{V}\left[\frac{1}{b}\sum_{i\in B}\boldsymbol{x}_i \,\middle|\, B\right]\right] \\
&= \mathbb{E}\left[\mathrm{tr}\mathbb{V}\left[\sum_{i\in B}\left(\frac{1}{b}\boldsymbol{x}_i\right) \,\middle|\, B\right]\right] \\
&\leq \rho\mathbb{V}S + \rho(\mathbb{E}S)^2 + (1-\rho)\,\mathbb{E}V, \quad (19)
\end{aligned}$$

where

$$\begin{aligned}
S &\triangleq \sum_{i\in B}\sqrt{\mathrm{tr}\mathbb{V}\left[\frac{1}{b}\boldsymbol{x}_i\right]} = \frac{1}{b}\sum_{i\in B}\sqrt{\mathrm{tr}\mathbb{V}[\boldsymbol{x}_i]}, \\
V &\triangleq \sum_{i\in B}\mathrm{tr}\mathbb{V}\left[\frac{1}{b}\boldsymbol{x}_i\right] = \frac{1}{b^2}\sum_{i\in B}\mathrm{tr}\mathbb{V}[\boldsymbol{x}_i].
\end{aligned}$$

In our context, S is the batch average of the standard deviations, and V is the batch average of the variance (scaled with a factor of $1/b$).

Notice that S is an b-sample average of the standard deviations. Therefore, if π is an unbiased subsampling strategy, we retrieve the population average standard deviation as

$$\mathbb{E}_{B\sim\pi}[S] = \mathbb{E}_{B\sim\pi}\left[\frac{1}{b}\sum_{i\in B}\sqrt{\mathrm{tr}\mathbb{V}[\boldsymbol{x}_i]}\right] = \frac{1}{n}\sum_{i=1}^n \sqrt{\mathrm{tr}\mathbb{V}[\boldsymbol{x}_i]}. \quad (20)$$

Under a similar reasoning, the variance of the standard deviations follows as

$$\begin{aligned}
&\mathbb{V}_{B\sim\pi}[S] \\
&= \mathbb{V}_{B\sim\pi}\left[\frac{1}{b}\sum_{i\in B}\sqrt{\mathrm{tr}\mathbb{V}[\boldsymbol{x}_i]}\right]
\end{aligned}$$

$$= \frac{1}{b_{\text{eff}}} \mathbb{V}_{i\sim\text{Uniform}\{1,\dots,n\}} \left[\sqrt{\text{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]}\right]$$
$$= \frac{1}{b_{\text{eff}}} \left(\frac{1}{n}\sum_{i=1}^{n} \text{tr}\mathbb{V}\left[\boldsymbol{x}_i\right] - \left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\text{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]}\right)^2\right), \quad (21)$$

where the last identity is the well-known formula for the variance: $\mathbb{V}X = \mathbb{E}X^2 - (\mathbb{E}X)^2$. Likewise, the average variance follows as

$$\mathbb{E}_{B\sim\pi} V = \frac{1}{b^2}\mathbb{E}_{B\sim\pi}\left[\sum_{i\in B} \text{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]\right]$$
$$= \frac{1}{b}\mathbb{E}_{B\sim\pi}\left[\frac{1}{b}\sum_{i\in B} \text{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]\right]$$
$$= \frac{1}{b}\left(\frac{1}{n}\sum_{i=1}^{n} \text{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]\right) \quad (22)$$

Plugging Eqs. (20) to (22) into Eq. (19), we have

$$\mathbb{E}_{B\sim\pi}\left[\text{tr}\mathbb{V}\left[\boldsymbol{x}_B \mid B\right]\right]$$
$$\leq \rho\mathbb{V}S + \rho(\mathbb{E}S)^2 + (1-\rho)\,\mathbb{E}V$$
$$= \frac{\rho}{b_{\text{eff}}}\left(\frac{1}{n}\sum_{i=1}^{n} \text{tr}\mathbb{V}\left[\boldsymbol{x}_i\right] - \left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\text{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]}\right)^2\right)$$
$$+ \rho\left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\text{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]}\right)^2$$
$$+ \frac{1-\rho}{b}\left(\frac{1}{n}\sum_{i=1}^{n} \text{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]\right)$$
$$= \left(\frac{\rho}{b_{\text{eff}}} + \frac{1-\rho}{b}\right)\left(\frac{1}{n}\sum_{i=1}^{n} \text{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]\right)$$
$$+ \rho\left(1 - \frac{1}{b_{\text{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\text{tr}\mathbb{V}\left[\boldsymbol{x}_i\right]}\right)^2. \quad (23)$$

Subsampling Variance The subsampling noise is straightforward. For this, we will denote the minibatch subsampling estimator of the component means as

$$\bar{\boldsymbol{x}}_B \triangleq \frac{1}{b}\sum_{i\in B}\bar{\boldsymbol{x}}_i.$$

Since each component estimator $\boldsymbol{x}_i$ is unbiased, the expectation conditional on the minibatch B is

$$\mathbb{E}\left[\bar{\boldsymbol{x}}_B \mid B\right] = \frac{1}{b}\sum_{i\in B}\bar{\boldsymbol{x}}_i.$$

Therefore,

$$\text{tr}\mathbb{V}_{B\sim\pi}\left[\mathbb{E}\left[\boldsymbol{x}_B \mid B\right]\right] = \text{tr}\mathbb{V}_{B\sim\pi}\left[\bar{\boldsymbol{x}}_B\right]$$
$$= \frac{1}{b_{\text{eff}}}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}\right\|_2^2\right). \quad (24)$$

Combining Eqs. (23) and (24) into Eq. (18) yields the result. Notice that the only inequality we used is Eq. (19), Lemma 9, in which equality holds if the equality in Assumption 3 holds. □

B.4. Doubly Stochastic Gradients

B.4.1. Expected Residual Condition (Theorem 2)

Theorem 2. *Let Assumption 4 to 6 hold. Then, we have:*

(i) *If* $(\mathrm{A}^{\mathrm{CVX}})$ *or* $(\mathrm{A}^{\mathrm{ITP}})$ *hold,* $\boldsymbol{g}_B$ *satisfies* $\mathrm{ER}(\mathcal{L}_{\mathrm{A}})$.

(ii) *If* (B) *holds,* $\boldsymbol{g}_B$ *satisfies* $\mathrm{ER}(\mathcal{L}_{\mathrm{B}})$.

where $\mathcal{L}_{\max} = \max\{\mathcal{L}_1, \dots, \mathcal{L}_n\}$,

$$\mathcal{L}_{\mathrm{A}} = \left(\frac{\rho}{b_{\mathrm{eff}}} + \frac{1-\rho}{b}\right)\mathcal{L}_{\max} + \rho\left(1 - \frac{1}{b_{\mathrm{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^n \mathcal{L}_i\right) + \frac{\mathcal{L}_{\mathrm{sub}}}{b_{\mathrm{eff}}}$$

$$\mathcal{L}_{\mathrm{B}} = \left(\frac{\rho}{b_{\mathrm{eff}}} + \frac{1-\rho}{b}\right)\left(\frac{1}{n}\sum_{i=1}^n \mathcal{L}_i\right) + \rho\left(1 - \frac{1}{b_{\mathrm{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^n \sqrt{\mathcal{L}_i}\right)^2 + \frac{\mathcal{L}_{\mathrm{sub}}}{b_{\mathrm{eff}}}.$$

Proof. From Theorem 1, we have

$$\begin{aligned}
&\mathrm{tr}\mathbb{V}\left[\boldsymbol{g}_B(\boldsymbol{x}) - \boldsymbol{g}_B(\boldsymbol{x}_*)\right] \\
&\le \left(\frac{\rho}{b_{\mathrm{eff}}} + \frac{1-\rho}{b}\right)\left(\frac{1}{n}\sum_{i=1}^n \mathrm{tr}\mathbb{V}\left[\boldsymbol{g}_i(\boldsymbol{x}) - \boldsymbol{g}_i(\boldsymbol{x}_*)\right]\right) \\
&\quad + \rho\left(1 - \frac{1}{b_{\mathrm{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^n \sqrt{\mathrm{tr}\mathbb{V}\left[\boldsymbol{g}_i(\boldsymbol{x}) - \boldsymbol{g}_i(\boldsymbol{x}_*)\right]}\right)^2 \\
&\quad + \frac{1}{b_{\mathrm{eff}}}\mathrm{tr}\mathbb{V}\left[\nabla f_B(\boldsymbol{x}) - \nabla F(\boldsymbol{x})\right],
\end{aligned}$$

where Assumption 5 yields

$$\begin{aligned}
&\le \left(\frac{\rho}{b_{\mathrm{eff}}} + \frac{1-\rho}{b}\right)\underbrace{\left(\frac{1}{n}\sum_{i=1}^n \mathrm{tr}\mathbb{V}\left[\boldsymbol{g}_i(\boldsymbol{x}) - \boldsymbol{g}_i(\boldsymbol{x}_*)\right]\right)}_{\triangleq T_{\mathrm{var}}} \\
&\quad + \rho\left(1 - \frac{1}{b_{\mathrm{eff}}}\right)\underbrace{\left(\frac{1}{n}\sum_{i=1}^n \sqrt{\mathrm{tr}\mathbb{V}\left[\boldsymbol{g}_i(\boldsymbol{x}) - \boldsymbol{g}_i(\boldsymbol{x}_*)\right]}\right)^2}_{\triangleq T_{\mathrm{cov}}} \\
&\quad + \frac{2\mathcal{L}_{\mathrm{sub}}}{b_{\mathrm{eff}}}\left(F(\boldsymbol{x}) - F(\boldsymbol{x}_*)\right) \\
&= \left(\frac{\rho}{b_{\mathrm{eff}}} + \frac{1-\rho}{b}\right)T_{\mathrm{var}} + \rho\left(1 - \frac{1}{b_{\mathrm{eff}}}\right)T_{\mathrm{cov}} \\
&\quad + \frac{2\mathcal{L}_{\mathrm{sub}}}{b_{\mathrm{eff}}}\left(F(\boldsymbol{x}) - F(\boldsymbol{x}_*)\right). \qquad (25)
\end{aligned}$$

Proof of (i) with $(\mathrm{A}^{\mathrm{CVX}})$ Since Assumption 6 $(\mathrm{A}^{\mathrm{CVX}})$ requires $f_1, \dots, f_n$ to be convex, F is also convex. Therefore, we can use the identity in Lemma 6 and

$$\mathrm{D}_F(\boldsymbol{x}, \boldsymbol{x}_*) = F(\boldsymbol{x}) - F(\boldsymbol{x}_*).$$

With that said, under $(\mathrm{A}^{\mathrm{CVX}})$, we have

$$\begin{aligned}
T_{\mathrm{var}} &\le \frac{1}{n}\sum_{i=1}^n \mathrm{tr}\mathbb{V}\left[\boldsymbol{g}_i(\boldsymbol{x}) - \boldsymbol{g}_i(\boldsymbol{x}_*)\right] \\
&= \frac{1}{n}\sum_{i=1}^n 2\mathcal{L}_i \mathrm{D}_{f_i}(\boldsymbol{x}, \boldsymbol{x}_*),
\end{aligned}$$

applying $\mathcal{L}_{\max} \ge \mathcal{L}_i$ for all $i = 1, \dots, n$,

$$\le 2\mathcal{L}_{\max}\frac{1}{n}\sum_{i=1}^n \mathrm{D}_{f_i}(\boldsymbol{x}, \boldsymbol{x}_*)$$

and Lemma 6,

$$= 2\mathcal{L}_{\max}\mathrm{D}_F(\boldsymbol{x}, \boldsymbol{x}_*). \qquad (26)$$

For T_{cov}, since

$$T_{\mathrm{cov}} = \left(\frac{1}{n}\sum_{i=1}^n \sqrt{\mathrm{tr}\mathbb{V}\left[\boldsymbol{g}_i(\boldsymbol{x}) - \boldsymbol{g}_i(\boldsymbol{x}_*)\right]}\right)^2$$

is monotonic w.r.t. the variance, we can apply $(\mathrm{A}^{\mathrm{CVX}})$ as

$$\le \frac{2}{n^2}\left(\sum_{i=1}^n \sqrt{\mathcal{L}_i \mathrm{D}_{f_i}(\boldsymbol{x}, \boldsymbol{x}_*)}\right)^2.$$

Now, applying the Cauchy-Schwarz inequality yields

$$\begin{aligned}
&\le \frac{2}{n^2}\left(\sum_{i=1}^n \mathcal{L}_i\right)\left(\sum_{i=1}^n \mathrm{D}_{f_i}(\boldsymbol{x}, \boldsymbol{x}_*)\right) \\
&= 2\left(\frac{1}{n}\sum_{i=1}^n \mathcal{L}_i\right)\left(\frac{1}{n}\sum_{i=1}^n \mathrm{D}_{f_i}(\boldsymbol{x}, \boldsymbol{x}_*)\right)
\end{aligned}$$

and by Lemma 6,

$$= 2\left(\frac{1}{n}\sum_{i=1}^n \mathcal{L}_i\right)\mathrm{D}_F(\boldsymbol{x}, \boldsymbol{x}_*). \qquad (27)$$

Plugging Eqs. (26) and (27) into Eq. (25), we have

$$\begin{aligned}
&\mathrm{tr}\mathbb{V}\left[\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x}_*)\right] \\
&\le \left(\frac{\rho}{b_{\mathrm{eff}}} + \frac{1-\rho}{b}\right)T_{\mathrm{var}} + \rho\left(1 - \frac{1}{b_{\mathrm{eff}}}\right)T_{\mathrm{cov}} \\
&\quad + \frac{2\mathcal{L}_{\mathrm{sub}}}{b_{\mathrm{eff}}}\left(F(\boldsymbol{x}) - F(\boldsymbol{x}_*)\right) \\
&\le \left(\frac{\rho}{b_{\mathrm{eff}}} + \frac{1-\rho}{b}\right)2\mathcal{L}_{\max}\mathrm{D}_F(\boldsymbol{x}, \boldsymbol{x}_*) \\
&\quad + \rho\left(1 - \frac{1}{b_{\mathrm{eff}}}\right)2\left(\frac{1}{n}\sum_{i=1}^n \mathcal{L}_i\right)\mathrm{D}_F(\boldsymbol{x}, \boldsymbol{x}_*) \\
&\quad + \frac{1}{b_{\mathrm{eff}}}2\mathcal{L}_{\mathrm{sub}}\mathrm{D}_F(\boldsymbol{x}, \boldsymbol{x}_*). \\
&= 2\left(\left(\frac{\rho}{b_{\mathrm{eff}}} + \frac{1-\rho}{b}\right)\mathcal{L}_{\max} + \rho\left(1 - \frac{1}{b_{\mathrm{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^n \mathcal{L}_i\right)\right. \\
&\quad \left. + \frac{1}{b_{\mathrm{eff}}}\mathcal{L}_{\mathrm{sub}}\right)\mathrm{D}_F(\boldsymbol{x}, \boldsymbol{x}_*) \\
&= 2\left(\left(\frac{\rho}{b_{\mathrm{eff}}} + \frac{1-\rho}{b}\right)\mathcal{L}_{\max} + \rho\left(1 - \frac{1}{b_{\mathrm{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^n \mathcal{L}_i\right)\right.
\end{aligned}$$

$$+\frac{1}{b_{\text{eff}}}\mathcal{L}_{\text{sub}}\Bigg)\left(F\left(\boldsymbol{x}\right)-F\left(\boldsymbol{x}_*\right)\right).$$

Proof of (i) with (A^{ITP}) From Assumption 6 (A^{ITP}), we have

$$\begin{aligned}T_{\text{var}} &= \frac{1}{n}\sum_{i=1}^{n}\text{tr}\mathbb{V}\left[\boldsymbol{g}_i\left(\boldsymbol{x}\right)-\boldsymbol{g}_i\left(\boldsymbol{x}_*\right)\right]\\ &\leq \frac{1}{n}\sum_{i=1}^{n}2\mathcal{L}_i\left(f_i\left(\boldsymbol{x}\right)-f_i\left(\boldsymbol{x}_*\right)\right),\end{aligned}$$

applying $\mathcal{L}_{\max}\geq\mathcal{L}_i$ for all $i=1,\ldots,n$,

$$\begin{aligned}&\leq 2\mathcal{L}_{\max}\frac{1}{n}\sum_{i=1}^{n}\left(f_i\left(\boldsymbol{x}\right)-f_i\left(\boldsymbol{x}_*\right)\right)\\ &= 2\mathcal{L}_{\max}\left(F\left(\boldsymbol{x}\right)-F\left(\boldsymbol{x}_*\right)\right).\end{aligned}\tag{28}$$

Similarly,

$$T_{\text{cov}}=\left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\text{tr}\mathbb{V}\left[\boldsymbol{g}_i\left(\boldsymbol{x}\right)-\boldsymbol{g}_i\left(\boldsymbol{x}_*\right)\right]}\right)^2,$$

applying (A^{ITP}),

$$\leq\frac{2}{n^2}\left(\sum_{i=1}^{n}\sqrt{\mathcal{L}_i\left(f_i\left(\boldsymbol{x}\right)-f_i\left(\boldsymbol{x}_*\right)\right)}\right)^2,$$

and applying the Cauchy-Schwarz inequality,

$$\begin{aligned}&\leq\frac{2}{n^2}\left(\sum_{i=1}^{n}\mathcal{L}_i\right)\left(\sum_{i=1}^{n}f_i\left(\boldsymbol{x}\right)-f_i\left(\boldsymbol{x}_*\right)\right)\\ &=2\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}f_i\left(\boldsymbol{x}\right)-f_i\left(\boldsymbol{x}_*\right)\right)\\ &=2\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i\right)\left(F\left(\boldsymbol{x}\right)-F\left(\boldsymbol{x}_*\right)\right).\end{aligned}\tag{29}$$

Plugging Eqs. (28) and (29) into Eq. (25), we have

$$\begin{aligned}&\text{tr}\mathbb{V}\left[\boldsymbol{g}\left(\boldsymbol{x}\right)-\boldsymbol{g}\left(\boldsymbol{x}_*\right)\right]\\ &\leq\left(\frac{\rho}{b_{\text{eff}}}+\frac{1-\rho}{b}\right)T_{\text{var}}+\rho\left(1-\frac{1}{b_{\text{eff}}}\right)T_{\text{cov}}\\ &\quad+\frac{2\mathcal{L}_{\text{sub}}}{b_{\text{eff}}}\left(F\left(\boldsymbol{x}\right)-F\left(\boldsymbol{x}_*\right)\right)\\ &\leq\left(\frac{\rho}{b_{\text{eff}}}+\frac{1-\rho}{b}\right)2\mathcal{L}_{\max}\left(F\left(\boldsymbol{x}\right)F\left(\boldsymbol{x}_*\right)\right)\\ &\quad+\rho\left(1-\frac{1}{b_{\text{eff}}}\right)2\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i\right)\left(F\left(\boldsymbol{x}\right)-F\left(\boldsymbol{x}_*\right)\right)\\ &\quad+\frac{1}{b_{\text{eff}}}2\mathcal{L}_{\text{sub}}\left(F\left(\boldsymbol{x}\right)-F\left(\boldsymbol{x}_*\right)\right).\\ &=2\Bigg(\left(\frac{\rho}{b_{\text{eff}}}+\frac{1-\rho}{b}\right)\mathcal{L}_{\max}+\rho\left(1-\frac{1}{b_{\text{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i\right)\\ &\quad+\frac{1}{b_{\text{eff}}}\mathcal{L}_{\text{sub}}\Bigg)\left(F\left(\boldsymbol{x}\right)-F\left(\boldsymbol{x}_*\right)\right).\end{aligned}$$

Proof of (ii) From Assumption 6 (B), we have

$$\begin{aligned}T_{\text{var}} &= \frac{1}{n}\sum_{i=1}^{n}\text{tr}\mathbb{V}\left[\boldsymbol{g}_i\left(\boldsymbol{x}\right)-\boldsymbol{g}_i\left(\boldsymbol{x}_*\right)\right]\\ &\leq\frac{1}{n}\sum_{i=1}^{n}2\mathcal{L}_i\left(F\left(\boldsymbol{x}\right)-F\left(\boldsymbol{x}_*\right)\right)\\ &=2\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i\right)\left(F\left(\boldsymbol{x}\right)-F\left(\boldsymbol{x}_*\right)\right).\end{aligned}\tag{30}$$

And,

$$\begin{aligned}T_{\text{cov}} &= \left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\text{tr}\mathbb{V}\left[\boldsymbol{g}_i\left(\boldsymbol{x}\right)-\boldsymbol{g}_i\left(\boldsymbol{x}_*\right)\right]}\right)^2\\ &\leq\frac{2}{n^2}\left(\sum_{i=1}^{n}\sqrt{\mathcal{L}_i\left(F\left(\boldsymbol{x}\right)-F\left(\boldsymbol{x}_*\right)\right)}\right)^2\\ &=2\left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\mathcal{L}_i}\right)^2\left(F\left(\boldsymbol{x}\right)-F\left(\boldsymbol{x}_*\right)\right).\end{aligned}\tag{31}$$

Plugging Eqs. (30) and (31) into Eq. (25), we have

$$\begin{aligned}&\text{tr}\mathbb{V}\left[\boldsymbol{g}\left(\boldsymbol{x}\right)-\boldsymbol{g}\left(\boldsymbol{x}_*\right)\right]\\ &\leq\left(\frac{\rho}{b_{\text{eff}}}+\frac{1-\rho}{b}\right)2\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i\right)\left(F\left(\boldsymbol{x}\right)-F\left(\boldsymbol{x}_*\right)\right)\\ &\quad+\rho\left(1-\frac{1}{b_{\text{eff}}}\right)2\left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\mathcal{L}_i}\right)^2\left(F\left(\boldsymbol{x}\right)-F\left(\boldsymbol{x}_*\right)\right)\\ &\quad+\frac{1}{b_{\text{eff}}}2\mathcal{L}_{\text{sub}}\left(F\left(\boldsymbol{x}\right)-F\left(\boldsymbol{x}_*\right)\right),\\ &=2\Bigg(\left(\frac{\rho}{b_{\text{eff}}}+\frac{1-\rho}{b}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i\right)\\ &\quad+\rho\left(1-\frac{1}{b_{\text{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\mathcal{L}_i}\right)^2\\ &\quad+\frac{1}{b_{\text{eff}}}\mathcal{L}_{\text{sub}}\Bigg)\left(F\left(\boldsymbol{x}\right)-F\left(\boldsymbol{x}_*\right)\right).\end{aligned}$$

□

B.4.2. Bounded Variance Condition (Theorem 3)

Theorem 3. *Let Assumption 4 and 7 hold. Then, $\boldsymbol{g}_B$ satisfies* BV (σ^2)*, where*

$$\sigma^2 = \left(\frac{\rho}{b_{\text{eff}}} + \frac{1-\rho}{b}\right)\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n \sigma_i^2\right) + \rho\left(1 - \frac{1}{b_{\text{eff}}}\right)\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n \sigma_i\right)^2 + \frac{\tau^2}{b_{\text{eff}}}.$$

Equality in Definition 2 holds if equality in Assumption 4 holds.

Proof. For any element of the solution set $\boldsymbol{x}_* = \arg\min_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$, by Theorem 1, we have

$$\begin{aligned}\mathrm{tr}\mathbb{V}\left[\boldsymbol{g}_B(\boldsymbol{x}_*)\right] &\le \left(\frac{\rho}{b_{\text{eff}}} + \frac{1-\rho}{b}\right)\left(\frac{1}{n}\sum_{i=1}^n \mathrm{tr}\mathbb{V}\left[\boldsymbol{g}_i(\boldsymbol{x}_*)\right]\right) \\ &\quad + \rho\left(1 - \frac{1}{b_{\text{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^n \sqrt{\mathrm{tr}\mathbb{V}\left[\boldsymbol{g}_i(\boldsymbol{x}_*)\right]}\right)^2 \\ &\quad + \frac{1}{b_{\text{eff}}}\mathrm{tr}\mathbb{V}\left[\nabla f_B(\boldsymbol{x}_*)\right].\end{aligned}$$

Applying Assumption 7, we have

$$\begin{aligned}&\le \left(\frac{\rho}{b_{\text{eff}}} + \frac{1-\rho}{b}\right)\left(\frac{1}{n}\sum_{i=1}^n \sigma_i^2\right) \\ &\quad + \left(1 - \frac{1}{b_{\text{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^n \sqrt{\sigma_i^2}\right)^2 \\ &\quad + \frac{1}{b_{\text{eff}}}\tau^2 \\ &= \left(1 - \frac{1}{b_{\text{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^n \sigma_i^2\right) \\ &\quad + \rho\left(1 - \frac{1}{b_{\text{eff}}}\right)\left(\frac{1}{n}\sum_{i=1}^n \sigma_i\right)^2 \\ &\quad + \frac{1}{b_{\text{eff}}}\tau^2.\end{aligned}$$

□

B.4.3. Complexity Analysis (Corollary 2)

Lemma 10. *Let the objective function F satisfy Assumption 2, π be sampling b samples without replacement, and all elements of the solution set $\boldsymbol{x}_* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$ be stationary points of F. Then, the subsampling estimator ∇f_B satisfies the* ER *condition as*

$$\begin{aligned}&\mathrm{tr}\mathbb{V}_{B\sim\pi}\left[\nabla f_B(\boldsymbol{x}) - \nabla f_B(\boldsymbol{x}_*)\right] \\ &\quad \le 2\frac{n-b}{b(n-1)} L_{\max}\left(F(\boldsymbol{x}) - F(\boldsymbol{x}_*)\right),\end{aligned}$$

where $L_{\max} = \max\{L_1, \dots, L_n\}$.

Proof. Consider that, for any random vector $\boldsymbol{x}$,

$$\mathrm{tr}\mathbb{V}\left[\boldsymbol{x}^2\right] \le \mathbb{E}\|\boldsymbol{x}\|_2^2$$

holds. Also, sampling without replacement achieves $b_{\text{eff}} = \frac{(n-1)b}{n-b}$. Therefore, we have

$$\begin{aligned}&\mathrm{tr}\mathbb{V}_{B\sim\pi}\left[\nabla f_B(\boldsymbol{x}) - \nabla f_B(\boldsymbol{x}_*)\right] \\ &= \frac{n-b}{b(n-1)}\mathrm{tr}\mathbb{V}\left[\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}_*)\right] \\ &\le \frac{n-b}{b(n-1)}\left(\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}_*)\|_2^2\right),\end{aligned}$$

and from Assumption 2,

$$= \frac{n-b}{b(n-1)}\left(\frac{1}{n}\sum_{i=1}^n 2L_i \mathrm{D}_{f_i}(\boldsymbol{x}, \boldsymbol{x}_*)\right).$$

Using the bound $L_{\max} \ge L_i$ for all $i = 1, \dots, n$,

$$\le 2L_{\max}\frac{n-b}{b(n-1)}\left(\frac{1}{n}\sum_{i=1}^n \mathrm{D}_{f_i}(\boldsymbol{x}, \boldsymbol{x}_*)\right),$$

applying Lemma 6,

$$= 2L_{\max}\frac{n-b}{b(n-1)}\mathrm{D}_F(\boldsymbol{x}, \boldsymbol{x}_*),$$

and since $\boldsymbol{x}_*$ is a stationary point of F,

$$= 2\frac{n-b}{b(n-1)} L_{\max}\left(F(\boldsymbol{x}) - F(\boldsymbol{x}_*)\right).$$

□

Corollary 2. *Let the objective F satisfy Assumption 1 and 2, the global optimum $\boldsymbol{x}_* = \arg\min_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$ be a stationary point of F, the component gradient estimators $\boldsymbol{g}_1, \dots, \boldsymbol{g}_n$ satisfy Assumption 6 (B) and 7, and π be b-minibatch sampling without replacement. Then the last iterate of SGD with $\boldsymbol{g}_B$ is ϵ-close to $\boldsymbol{x}_*$ as $\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}_*\|_2^2 \leq \epsilon$ after a number of iterations of at least*

$$T \geq 2\max\left(C_{\text{var}}\frac{1}{\epsilon},\ C_{\text{bias}}\right)\log\left(2\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2\frac{1}{\epsilon}\right)$$

for some fixed stepsize where

$$C_{\text{var}} = \frac{2}{b}\left(\frac{1}{n}\sum_{i=1}^n \frac{\sigma_i^2}{\mu^2}\right) + 2\left(\frac{1}{n}\sum_{i=1}^n \frac{\sigma_i}{\mu}\right)^2 + \frac{2}{b}\frac{\tau^2}{\mu^2},$$

$$C_{\text{bias}} = \frac{2}{b}\left(\frac{1}{n}\sum_{i=1}^n \frac{\mathcal{L}_i}{\mu}\right) + 2\left(\frac{1}{n}\sum_{i=1}^n \sqrt{\frac{\mathcal{L}_i}{\mu}}\right)^2 + \frac{2}{b}\frac{L}{\mu}.$$

Proof. From Assumption 2 and the assumption that $\boldsymbol{x}_*$ is a stationary point, Lemma 10 establishes that ∇f_B satisfies the ER $(\mathcal{L}_{\text{sub}})$ holds with

$$\mathcal{L}_{\text{sub}} = \frac{n-b}{(n-1)b}L_{\max}.$$

Therefore, Assumption 5 holds. Furthermore, since the component gradient estimators satisfy Assumption 6 (B) and Assumption 3 always hold with $\rho = 1$, we can apply Theorem 2 which estblishes that $\boldsymbol{g}_B$ satisfies ER $(\mathcal{L})$ with

$$\mathcal{L} = \frac{n-b}{(n-1)b}\left(\frac{1}{n}\sum_{i=1}^n \mathcal{L}_i\right) + \frac{n(b-1)}{(n-1)b}\left(\frac{1}{n}\sum_{i=1}^n \sqrt{\mathcal{L}_i}\right)^2 + \frac{n-b}{(n-1)b}L_{\max}.$$

Furthermore, under Assumption 7, Theorem 3 shows that BV (σ^2) holds with

$$\sigma^2 = \frac{n-b}{(n-1)b}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i^2\right) + \frac{n(b-1)}{(n-1)b}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i\right)^2 + \frac{n-b}{(n-1)b}\tau^2.$$

Since both ER $(\mathcal{L})$ and BV (σ^2) hold and F satisfies Assumption 1, we can now invoke Lemma 1, which guarantees that we can obtain an ϵ-accurate solution after

$$T \geq 2\max\Bigg(\underbrace{\frac{\sigma^2}{\mu^2}}_{C_{\text{var}}}\ \frac{1}{\epsilon},\ \underbrace{\frac{\mathcal{L}+L}{\mu}}_{C_{\text{bias}}}\Bigg)\log\left(2\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2\frac{1}{\epsilon}\right)$$

iterations and fixed stepsize of

$$\gamma = \min\left(\frac{\epsilon\mu}{2\sigma^2}, \frac{1}{2(\mathcal{L}+L)}\right).$$

The constants in the lower bound on the number of required iterations can be made more precise as

$$C_{\text{var}} = \frac{n-b}{(n-1)b}\left(\frac{1}{n}\sum_{i=1}^n \frac{\sigma_i^2}{\mu^2}\right) + \frac{n(b-1)}{(n-1)b}\left(\frac{1}{n}\sum_{i=1}^n \frac{\sigma_i}{\mu}\right)^2 + \frac{n-b}{(n-1)b}\frac{\tau^2}{\mu^2}$$

$$C_{\text{bias}} = \frac{n-b}{(n-1)b}\left(\frac{1}{n}\sum_{i=1}^n \frac{\mathcal{L}_i}{\mu}\right) + \frac{n(b-1)}{(n-1)b}\left(\frac{1}{n}\sum_{i=1}^n \sqrt{\frac{\mathcal{L}_i}{\mu}}\right)^2 + \frac{n-b}{(n-1)b}\frac{L}{\mu}.$$

Using the fact that $(n-b)/n \leq (n-1)/n \leq 2$ for all $n \geq 2$ yields the simplified constants in the statement. □

B.5. Random Reshuffling of Stochastic Gradients

B.5.1. Gradient Variance Conditions (Lemma 11, Lemma 12)

Lemma 11. *Let the objective function satisfy Assumption 2, B be any b-minibatch of indices such that $B \subseteq \{1, \dots, n\}$ and the component gradient estimators $\boldsymbol{g}_1, \dots, \boldsymbol{g}_n$ satisfy Assumption 6 ($\mathrm{A}^{\mathrm{CVX}}$). Then, $\boldsymbol{g}_B$ is convex-smooth in expectation such that*

$$\mathbb{E}_\varphi \|\boldsymbol{g}_B(\boldsymbol{x}) - \boldsymbol{g}_B(\boldsymbol{x}_*)\|_2^2 \leq 2(\mathcal{L}_{\max} + L_{\max}) \mathrm{D}_{f_B}(\boldsymbol{x}, \boldsymbol{x}_*),$$

for any $\boldsymbol{x} \in \mathcal{X}$, where

$$\begin{aligned} \boldsymbol{x}_* &= \arg\min_{\boldsymbol{x} \in \mathcal{X}} F(\boldsymbol{x}), \\ \mathcal{L}_{\max} &= \max\{\mathcal{L}_1, \dots, \mathcal{L}_n\}, \\ L_{\max} &= \max\{L_1, \dots, L_n\}. \end{aligned}$$

Proof. Notice that, for this Lemma, we do not assume that the minibatch B is a random variable. Therefore, the only randomness is the stochasticity of the component gradient estimators $\boldsymbol{g}_1, \dots, \boldsymbol{g}_n$.

Now, from the property of the variance, we can decompose the expected squared norm as

$$\begin{aligned} &\mathbb{E}\|\boldsymbol{g}_B(\boldsymbol{x}) - \boldsymbol{g}_B(\boldsymbol{x}_*)\|_2^2 \\ &= \underbrace{\mathrm{tr}\mathbb{V}_\varphi[\boldsymbol{g}_B(\boldsymbol{x}) - \boldsymbol{g}_B(\boldsymbol{x}_*)]}_{V_{\mathrm{com}}} + \underbrace{\|\nabla f_B(\boldsymbol{x}) - \nabla f_B(\boldsymbol{x}_*)\|_2^2}_{V_{\mathrm{sub}}}. \end{aligned}$$

First, the contribution of the variances of the component gradient estimators follows as

$$\begin{aligned} V_{\mathrm{com}} &= \mathrm{tr}\mathbb{V}_\varphi[\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x}_*)] \\ &= \mathrm{tr}\mathbb{V}_\varphi\left[\frac{1}{b}\sum_{i \in B} \boldsymbol{g}_i(\boldsymbol{x}) - \boldsymbol{g}_i(\boldsymbol{x}_*)\right], \end{aligned}$$

applying Eq. (14) of Lemma 3,

$$\leq \frac{1}{b}\sum_{i \in B} \mathrm{tr}\mathbb{V}_\varphi[\boldsymbol{g}_i(\boldsymbol{x}) - \boldsymbol{g}_i(\boldsymbol{x}_*)], \tag{32}$$

and then Assumption 6 ($\mathrm{A}^{\mathrm{CVX}}$),

$$\leq \frac{1}{b}\sum_{i \in B} 2\mathcal{L}_i \mathrm{D}_{f_i}(\boldsymbol{x}, \boldsymbol{x}_*).$$

Now, since $\mathcal{L}_{\max} \geq \mathcal{L}_i$ for all $i = 1, \dots, n$,

$$\begin{aligned} &\leq 2\mathcal{L}_{\max}\frac{1}{b}\sum_{i \in B} \mathrm{D}_{f_i}(\boldsymbol{x}, \boldsymbol{x}_*) \\ &= 2\mathcal{L}_{\max}\mathrm{D}_{f_B}(\boldsymbol{x}, \boldsymbol{x}_*). \end{aligned}$$

On the other hand, the squared error of subsampling (it is not the variance since we do not take expectation over the batches) follows as

$$\begin{aligned} V_{\mathrm{sub}} &= \|\nabla f_B(\boldsymbol{x}) - \nabla f_B(\boldsymbol{x}_*)\|_2^2 \\ &= \left\|\frac{1}{b}\sum_{i \in B} \nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}_*)\right\|_2^2, \end{aligned}$$

by Jensen's inequality,

$$\leq \frac{1}{b}\sum_{i \in B} \|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}_*)\|_2^2,$$

from Assumption 2,

$$\leq \frac{1}{b}\sum_{i \in B} 2L_i \mathrm{D}_{f_i}(\boldsymbol{x}, \boldsymbol{x}_*)$$

and since $L_{\max} \geq L_i$ for all $i = 1, \dots, n$,

$$\begin{aligned} &\leq 2L_{\max}\frac{1}{b}\sum_{i \in B} \mathrm{D}_{f_i}(\boldsymbol{x}, \boldsymbol{x}_*) \\ &= 2L_{\max}\mathrm{D}_{f_B}(\boldsymbol{x}, \boldsymbol{x}_*). \end{aligned}$$

Combining the bound on V_{com} and V_{sub} immediately yields the result. □

Lemma 12. *For any b-minibatch reshuffling strategy, the squared error of the reference point of the Lyapunov function (Eq. (10)) under reshuffling is bounded as*

$$\mathbb{E}\|\boldsymbol{x}_*^i - \boldsymbol{x}_*\|_2^2 \leq \frac{\gamma^2 n}{4b^2}\tau^2$$

for all $i = 1, \dots, p$, *where* $\boldsymbol{x}_* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$.

Proof. The proof is a generalization of Mishchenko et al. (2020, Proposition 1), where we sample b-minibatches instead of single datapoints. Recall that P denotes the (possibly random) partitioning of the n datapoints into b-minibatches $P_1, \dots, P_p$. From the definition of the squared error of the Lyapunov function in Eq. (10), we have

$$\mathbb{E}\left[\|\boldsymbol{x}_*^i - \boldsymbol{x}_*\|_2^2\right] = \mathbb{E}\left[\left\|\Pi_{\mathcal{X}}\left(\boldsymbol{x}_* - \sum_{k=0}^{i-1}\gamma\nabla f_{P_i}(\boldsymbol{x}_*)\right) - \Pi_{\mathcal{X}}(\boldsymbol{x}_*)\right\|_2^2\right],$$

and since the projection onto a convex set under a Euclidean metric is non-expansive,

$$\leq \mathbb{E}\left[\left\|\boldsymbol{x}_* - \sum_{k=0}^{i-1}\gamma\nabla f_{P_i}(\boldsymbol{x}_*) - \boldsymbol{x}_*\right\|_2^2\right] = \mathbb{E}\left[\left\|\sum_{k=0}^{i-1}\gamma\nabla f_{P_i}(\boldsymbol{x}_*)\right\|_2^2\right],$$

introducing a factor of i in and out of the squared norm,

$$= \frac{i^2}{2}\mathbb{E}\left[\left\|\frac{1}{i}\sum_{k=0}^{i-1}\gamma\nabla f_{P_i}(\boldsymbol{x}_*)\right\|_2^2\right] = \frac{\gamma^2 i^2}{2}\mathbb{E}\left[\left\|\frac{1}{i}\sum_{k=0}^{i-1}\nabla f_{P_i}(\boldsymbol{x}_*)\right\|_2^2\right].$$

Now notice that $\frac{1}{i}\sum_{j=0}^{i-1}\nabla f_{P_i}(\boldsymbol{x}_*)$ is a sample average of ib samples drawn without replacement. Therefore, it is an unbiased estimate of $\nabla F(\boldsymbol{x}_*)$. This implies

$$\mathbb{E}\left[\|\boldsymbol{x}_*^i - \boldsymbol{x}_*\|_2^2\right] = \frac{\gamma^2 i^2}{2}\mathbb{E}\left[\left\|\frac{1}{i}\sum_{k=0}^{i-1}\nabla f_{P_i}(\boldsymbol{x}_*)\right\|_2^2\right] = \frac{\gamma^2 i^2}{2}\operatorname{tr}\mathbb{V}\left[\frac{1}{i}\sum_{k=0}^{i-1}\nabla f_{P_i}(\boldsymbol{x}_*)\right],$$

and from Lemma 2 with a sample size of ib,

$$= \frac{\gamma^2 i^2}{2}\frac{n-ib}{(n-1)ib}\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\boldsymbol{x}_*)\|_2^2 = \frac{\gamma^2 i\left(\frac{n}{b}-i\right)}{2(n-1)}\tau^2.$$

Notice that this is a quadratic with respect to i, where the maximum is obtained by $i = n/2b$. Then,

$$\mathbb{E}\left[\|\boldsymbol{x}_*^i - \boldsymbol{x}_*\|_2^2\right] \leq \frac{\gamma^2\left(\frac{n}{2b}\right)^2}{2(n-1)}\tau^2 = \frac{\gamma^2 n^2}{8b^2(n-1)}\tau^2,$$

and using the bound $n/(n-1) \leq 2$ for all $n \geq 2$,

$$\leq \frac{\gamma^2 n}{4b^2}\tau^2.$$

□

B.5.2. Convergence Analysis (Theorem 5)

Theorem 5. *Let the objective F satisfy Assumption 1 and 2, where, each component f_i is additionally μ-strongly convex and Assumption 6 ($\mathrm{A}^{\mathrm{CVX}}$), 7 hold. Then, the last iterate $\boldsymbol{x}_T$ of doubly SGD-RR with a stepsize satisfying $\gamma < 1/(\mathcal{L}_{\max} + L_{\max})$ guarantees*

$$\mathbb{E}\|\boldsymbol{x}_{K+1}^0 - \boldsymbol{x}_*\|_2^2 \leq r^{Kp}\|\boldsymbol{x}_1^0 - \boldsymbol{x}_*\|_2^2 + C_{\mathrm{var}}^{\mathrm{sub}}\gamma^2 + C_{\mathrm{var}}^{\mathrm{com}}\gamma$$

where $p = n/b$ is the number of epochs, $\boldsymbol{x}_* = \arg\min_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$, $r = 1 - \gamma\mu$ is the contraction coefficient,

$$C_{\mathrm{var}}^{\mathrm{com}} = \frac{4}{\mu b}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i^2\right) + \frac{4}{\mu}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i\right)^2, \text{ and}$$

$$C_{\mathrm{var}}^{\mathrm{sub}} = \frac{1}{4}\frac{L_{\max}}{\mu}\frac{n}{b^2}\left(\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}_*)\|_2^2\right).$$

Proof. The key element of the analysis of random reshuffling is that the Lyapunov function that achieves a fast convergence is $\|\boldsymbol{x}_k^{i+1} - \boldsymbol{x}_*^{i+1}\|_2^2$ not $\|\boldsymbol{x}_k^{i+1} - \boldsymbol{x}_*\|_2^2$. This stems from the well-known fact that random reshuffling results in a conditionally biased gradient estimator.

Recall that P denotes the partitioning of the n datapoints into b-minibatches $P_1, \ldots, P_p$. As usual, we first expand the Lyapunov function as

$$\begin{aligned}
&\|\boldsymbol{x}_k^{i+1} - \boldsymbol{x}_*^{i+1}\|_2^2 \\
&= \|\Pi_{\mathcal{X}}(\boldsymbol{x}_k^i - \gamma\, \boldsymbol{g}_{P_i}(\boldsymbol{x}_k^i)) - \Pi_{\mathcal{X}}(\boldsymbol{x}_*^i - \gamma\, \nabla f_{P_i}(\boldsymbol{x}_*))\|_2^2
\end{aligned}$$

and since the projection onto a convex set under a Euclidean metric is non-expansive,

$$\begin{aligned}
&\leq \|(\boldsymbol{x}_k^i - \gamma\, \boldsymbol{g}_{P_i}(\boldsymbol{x}_k^i)) - (\boldsymbol{x}_*^i - \gamma\, \nabla f_{P_i}(\boldsymbol{x}_*))\|_2^2 \\
&= \|\boldsymbol{x}_k^i - \boldsymbol{x}_*\|_2^2 - 2\gamma\left\langle \boldsymbol{x}_k^i - \boldsymbol{x}_*^i\,, \boldsymbol{g}_{P_i}(\boldsymbol{x}_k^i) - \nabla f_{P_i}(\boldsymbol{x}_*)\right\rangle \\
&\quad + \gamma^2\|\boldsymbol{g}_{P_i}(\boldsymbol{x}_k^i) - \nabla f_{P_i}(\boldsymbol{x}_*)\|_2^2.
\end{aligned}$$

Taking expectation over the Monte Carlo noise conditional on the partitioning P,

$$\begin{aligned}
&\mathbb{E}_\varphi\|\boldsymbol{x}_k^{i+1} - \boldsymbol{x}_*^{i+1}\|_2^2 \\
&= \|\boldsymbol{x}_k^i - \boldsymbol{x}_*^i\|_2^2 - 2\gamma\left\langle \boldsymbol{x}_k^i - \boldsymbol{x}_*^i\,, \mathbb{E}_\varphi[\boldsymbol{g}_{P_i}(\boldsymbol{x}_k^i)] - \nabla f_{P_i}(\boldsymbol{x}_*)\right\rangle \\
&\quad + \gamma^2\mathbb{E}_\varphi\|\boldsymbol{g}_{P_i}(\boldsymbol{x}_k^i) - \nabla f_{P_i}(\boldsymbol{x}_*)\| \\
&= \|\boldsymbol{x}_k^i - \boldsymbol{x}_*^i\|_2^2 - 2\gamma\left\langle \boldsymbol{x}_k^i - \boldsymbol{x}_*^i\,, \nabla f_{P_i}(\boldsymbol{x}_k^i) - \nabla f_{P_i}(\boldsymbol{x}_*)\right\rangle \\
&\quad + \gamma^2\,\mathbb{E}_\varphi\|\boldsymbol{g}_{P_i}(\boldsymbol{x}_k^i) - \nabla f_{P_i}(\boldsymbol{x}_*)\|_2^2.
\end{aligned}$$

From the three-point identity, we can more precisely characterize the effect of the conditional bias such that

$$\begin{aligned}
&\left\langle \boldsymbol{x}_k^i - \boldsymbol{x}_*^i\,, \nabla f_{P_i}(\boldsymbol{x}_k^i) - \nabla f_{P_i}(\boldsymbol{x}_*)\right\rangle \\
&= \mathrm{D}_{f_{P_i}}(\boldsymbol{x}_*^i, \boldsymbol{x}_k^i) + \mathrm{D}_{f_{P_i}}(\boldsymbol{x}_k^i, \boldsymbol{x}_*) - \mathrm{D}_{f_{P_i}}(\boldsymbol{x}_*^i, \boldsymbol{x}_*).
\end{aligned}$$

For the gradient noise,

$$\begin{aligned}
&\mathbb{E}_\varphi\|\boldsymbol{g}_{P_i}(\boldsymbol{x}_k^i) - \nabla f_{P_i}(\boldsymbol{x}_*)\|_2^2 \\
&= \mathbb{E}_\varphi\|\boldsymbol{g}_{P_i}(\boldsymbol{x}_k^i) - \boldsymbol{g}_{P_i}(\boldsymbol{x}_*) + \boldsymbol{g}_{P_i}(\boldsymbol{x}_*) - \nabla f_{P_i}(\boldsymbol{x}_*)\|_2^2 \\
&\leq 2\mathbb{E}_\varphi\|\boldsymbol{g}_{P_i}(\boldsymbol{x}_k^i) - \boldsymbol{g}_{P_i}(\boldsymbol{x}_*)\|_2^2 + 2\mathbb{E}_\varphi\|\boldsymbol{g}_{P_i}(\boldsymbol{x}_*) - \nabla f_{P_i}(\boldsymbol{x}_*)\|_2^2 \\
&= 2\,\mathbb{E}_\varphi\|\boldsymbol{g}_{P_i}(\boldsymbol{x}_k^i) - \boldsymbol{g}_{P_i}(\boldsymbol{x}_*)\|_2^2 + 2\,\mathrm{tr}\mathbb{V}_\varphi\left[\boldsymbol{g}_{P_i}(\boldsymbol{x}_*)\right],
\end{aligned}$$

and from Lemma 11,

$$\leq 4\,(\mathcal{L}_{\max} + L_{\max})\,\mathrm{D}_{f_{P_i}}(\boldsymbol{x}_k^i, \boldsymbol{x}_*) + 2\mathrm{tr}\mathbb{V}_\varphi\left[\boldsymbol{g}_{P_i}(\boldsymbol{x}_*)\right]$$

Notice the variance term $\mathrm{tr}\mathbb{V}_\varphi\left[\boldsymbol{g}_{P_i}(\boldsymbol{x}_*)\right]$. This quantifies the amount of deviation from the trajectory of singly stochastic random reshuffling. As such, it quantifies how slower we will be compared to its fast rate.

Now, we will denote the σ-algebra formed by the randomness and the iterates up to the ith step of the kth epoch as $\mathcal{F}_k^i$ such that $(\mathcal{F}_k^i)_{k\geq 1, i\geq 1}$ is a filtration. Then,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\eta}_k^i\sim\varphi}\left[\|\boldsymbol{x}_k^{i+1} - \boldsymbol{x}_*^{i+1}\|_2^2 \,\middle|\, \mathcal{F}_k^i\right] \\
&\leq \|\boldsymbol{x}_k^i - \boldsymbol{x}_*^i\|_2^2 \\
&\quad - 2\gamma\left(\mathrm{D}_{f_{P_i}}(\boldsymbol{x}_*^i, \boldsymbol{x}_k^i) + \mathrm{D}_{f_{P_i}}(\boldsymbol{x}_k^i, \boldsymbol{x}_*) - \mathrm{D}_{f_{P_i}}(\boldsymbol{x}_*^i, \boldsymbol{x}_*)\right) \\
&\quad + 4\gamma^2\,(\mathcal{L}_{\max} + L_{\max})\,\mathrm{D}_{f_{P_i}}(\boldsymbol{x}_k^i, \boldsymbol{x}_*) \\
&\quad + 2\gamma^2\mathrm{tr}\mathbb{V}_\varphi\left[\boldsymbol{g}_{P_i}(\boldsymbol{x}_*)\right].
\end{aligned}$$

Now, the μ-strong convexity of the component functions imply $\mathrm{D}_{f_{P_i}}\left(\boldsymbol{x}_*^i, \boldsymbol{x}_k^i\right) \leq \frac{\mu}{2}\|\boldsymbol{x}_k^i - \boldsymbol{x}_*^i\|_2^2$. Therefore,

$$\begin{aligned}
&\leq \|\boldsymbol{x}_k^i - \boldsymbol{x}_*^i\|_2^2 \\
&\quad - 2\gamma\left(\frac{\mu}{2}\|\boldsymbol{x}_k^i - \boldsymbol{x}_*^i\|_2^2 + \mathrm{D}_{f_{P_i}}(\boldsymbol{x}_k^i, \boldsymbol{x}_*) - \mathrm{D}_{f_{P_i}}(\boldsymbol{x}_*^i, \boldsymbol{x}_*)\right) \\
&\quad + 4\gamma^2\,(\mathcal{L}_{\max} + L_{\max})\,\mathrm{D}_{f_{P_i}}(\boldsymbol{x}_k^i, \boldsymbol{x}_*) \\
&\quad + 2\gamma^2\mathrm{tr}\mathbb{V}_\varphi\left[\boldsymbol{g}_{P_i}(\boldsymbol{x}_*)\right],
\end{aligned}$$

and reorganizing the terms,

$$\begin{aligned}
&= (1 - \gamma\mu)\,\|\boldsymbol{x}_k^i - \boldsymbol{x}_*^i\|_2^2 \\
&\quad - 2\gamma\,(1 - 2\gamma\,(\mathcal{L}_{\max} + L_{\max}))\,\mathrm{D}_{f_{P_i}}(\boldsymbol{x}_k^i, \boldsymbol{x}_*) \\
&\quad + 2\gamma\,\mathrm{D}_{f_{P_i}}(\boldsymbol{x}_*^i, \boldsymbol{x}_*) \\
&\quad + \gamma^2 2\mathrm{tr}\mathbb{V}_\varphi\left[\boldsymbol{g}_{P_i}(\boldsymbol{x}_*)\right].
\end{aligned}$$

Taking full expectation,

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{x}_k^{i+1} - \boldsymbol{x}_*^{i+1}\|_2^2 \\
&\leq (1-\gamma\mu)\,\mathbb{E}\|\boldsymbol{x}_k^i - \boldsymbol{x}_*^i\|_2^2 \\
&\quad - 2\gamma\,(1 - 2\gamma\,(\mathcal{L}_{\max} + L_{\max}))\,\mathbb{E}\left[\mathrm{D}_{f_{P_i}}\left(\boldsymbol{x}_k^n, \boldsymbol{x}_*\right)\right] \\
&\quad + 2\gamma\,\mathbb{E}\left[\mathrm{D}_{f_{P_i}}\left(\boldsymbol{x}_*^i, \boldsymbol{x}_*\right)\right] \\
&\quad + 2\gamma^2\mathbb{E}\left[\mathrm{tr}\mathbb{V}_\varphi\left[\boldsymbol{g}_{P_i}\left(\boldsymbol{x}_*\right)\right]\right],
\end{aligned}$$

and as long as $\gamma < 1/(2\,(\mathcal{L}_{\max} + L_{\max}))$

$$\begin{aligned}
&\leq (1-\gamma\mu)\,\mathbb{E}\|\boldsymbol{x}_k^i - \boldsymbol{x}_*^i\|_2^2 + 2\gamma\underbrace{\mathbb{E}\left[\mathrm{D}_{f_{P_i}}\left(\boldsymbol{x}_*^i, \boldsymbol{x}_*\right)\right]}_{T_{\text{err}}} \\
&\quad + 2\gamma^2\underbrace{\mathbb{E}\left[\mathrm{tr}\mathbb{V}_\varphi\left[\boldsymbol{g}_{P_i}\left(\boldsymbol{x}_*\right)\right]\right]}_{T_{\text{var}}}. \qquad (33)
\end{aligned}$$

Bounding T_{err} From the definition of the Bregman divergence and L-smoothness, for all $j = 1, \ldots, n$, notice that we have

$$\begin{aligned}
\mathrm{D}_{f_j}(\boldsymbol{y}, \boldsymbol{x}) &= f_j(\boldsymbol{y}) - f_j(\boldsymbol{x}) - \left\langle \nabla f_j(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\right\rangle \\
&\leq \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2. \qquad (34)
\end{aligned}$$

for all $(\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{X}^2$. Given this, the Lyapunov error term

$$\mathbb{E}\left[\mathrm{D}_{f_{P_i}}\left(\boldsymbol{x}_k^i, \boldsymbol{x}_*\right)\right] = \mathbb{E}\left[\frac{1}{b}\sum_{j\in P_i}\mathrm{D}_{f_j}\left(\boldsymbol{x}_k^i, \boldsymbol{x}_*\right)\right]$$

can be bounded using L-smoothness by Eq. (34),

$$\leq \mathbb{E}\left[\frac{1}{b}\sum_{j\in P_i}\frac{L_j}{2}\|\boldsymbol{x}_k^i - \boldsymbol{x}_*\|_2^2\right]$$

and $L_{\max} \geq L_i$ for all $i = 1, \ldots, n$,

$$\begin{aligned}
&\leq \frac{L_{\max}}{2}\mathbb{E}\left[\frac{1}{b}\sum_{j\in P_i}\|\boldsymbol{x}_k^i - \boldsymbol{x}_*\|_2^2\right] \\
&= \frac{L_{\max}}{2}\mathbb{E}\|\boldsymbol{x}_k^i - \boldsymbol{x}_*\|_2^2. \qquad (35)
\end{aligned}$$

The squared error $\|\boldsymbol{x}_k^i - \boldsymbol{x}_*\|_2^2$ is bounded in Lemma 12 as

$$\mathbb{E}\|\boldsymbol{x}_k^i - \boldsymbol{x}_*\|_2^2 \leq \epsilon_{\text{sfl}}^2 \triangleq \frac{\gamma^2 n}{4b^2}\tau^2 < \infty. \qquad (36)$$

Bounding T_{var} Now, let's take a look at the variance term. First, notice that, by the Law of Total Expectation,

$$\mathbb{E}\left[\mathrm{tr}\mathbb{V}_\varphi\left[\boldsymbol{g}_{P_i}\left(\boldsymbol{x}_*\right)\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\mathrm{tr}\mathbb{V}_\varphi\left[\boldsymbol{g}_{P_i}\left(\boldsymbol{x}_*\right)\right] \mid P\right]\right].$$

Here,

$$\mathbb{E}\left[\mathrm{tr}\mathbb{V}_\varphi\left[\boldsymbol{g}_{P_i}\left(\boldsymbol{x}_*\right)\right] \mid P\right]$$

is the variance from selecting b samples without replacement. We can thus apply Lemma 9 with $b_{\text{eff}} = \frac{(n-1)b}{n-b}$ such that

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{tr}\mathbb{V}_\varphi\left[\boldsymbol{g}_{P_i}\left(\boldsymbol{x}_*\right)\right] \mid P\right] \\
&\leq \frac{n-b}{(n-1)\,b}\left(\frac{1}{n}\sum_{j=1}^n \sigma_j^2\right) + \frac{n\,(b-1)}{(n-1)\,b}\left(\frac{1}{n}\sum_{j=1}^n \sigma_j\right)^2,
\end{aligned}$$

which we will denote as

$$= \sigma^2 \qquad (37)$$

for clarity. Also, notice that σ^2 no longer depends on the partitioning.

Per-step Recurrence Equation Applying Eqs. (35) and (37) to Eq. (33), we now have the recurrence equation

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{x}_k^{i+1} - \boldsymbol{x}_*^{i+1}\|_2^2 &\leq (1-\gamma\mu)\,\mathbb{E}\|\boldsymbol{x}_k^i - \boldsymbol{x}_*^i\|_2^2 \\
&\quad + L_{\max}\epsilon_{\text{sfl}}^2\gamma + 2\sigma^2\gamma^2.
\end{aligned}$$

Now that we have a contraction of the Lyapunov function $\mathbb{E}\|\boldsymbol{x}_k^{i+1} - \boldsymbol{x}_*^{i+1}\|_2^2$, it remains to convert this that the Lyapunov function bounds our objective $\mathbb{E}\|\boldsymbol{x}_k^{i+1} - \boldsymbol{x}_*\|_2^2$. This can be achieved by noticing that, at the end of each epoch, we have $\boldsymbol{x}_{k+1} - \boldsymbol{x}_* = \boldsymbol{x}_k^p - \boldsymbol{x}_*^p$, and equivalently, we have $\boldsymbol{x}_k - \boldsymbol{x}_* = \boldsymbol{x}_k^0 - \boldsymbol{x}_*^0$ at the beginning of the epoch. The fact that the relationship with the original objective is only guaranteed at the endpoints (beginning and end of the epoch) is related to the fact that the bias of random reshuffling starts increasing at the beginning of the epoch and starts decreasing near the end.

Per-Epoch Recurrence Equation Nevertheless, this implies that by simply unrolling the recursion as in the analysis of regular SGD, we obtain a per-epoch contraction of

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{x}_{k+1}^0 - \boldsymbol{x}_*\|_2^2 &\leq (1-\gamma\mu)^p\,\mathbb{E}\|\boldsymbol{x}_k^0 - \boldsymbol{x}_*\|_2^2 \\
&\quad + \left(L_{\max}\epsilon_{\text{sfl}}^2\gamma + 2\sigma^2\gamma^2\right)\left(\sum_{i=0}^{p-1}(1-\mu\gamma)^i\right).
\end{aligned}$$

And after K epochs,

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{x}_{K+1}^0 - \boldsymbol{x}_*\|_2^2 \leq (1-\gamma\mu)^{pK}\,\mathbb{E}\|\boldsymbol{x}_0^0 - \boldsymbol{x}_*\|_2^2 \\
&+ \left(L_{\max}\epsilon_{\text{sfl}}^2\gamma + 2\sigma^2\gamma^2\right)\left(\sum_{i=0}^{p-1}(1-\mu\gamma)^i\right)\left(\sum_{j=0}^{pK-1}(1-\mu\gamma)^{pj}\right).
\end{aligned}$$

Note that $T = pK$.

As done by Mishchenko et al. (2020), the product of sums can be bounded as

$$\begin{aligned}
&\left(\sum_{i=0}^{p-1}(1-\mu\gamma)^i\right)\left(\sum_{j=0}^{T-1}(1-\mu\gamma)^{pj}\right)\\
&=\sum_{i=0}^{p-1}\sum_{j=0}^{T-1}(1-\mu\gamma)^i(1-\mu\gamma)^{pj}\\
&=\sum_{i=0}^{p-1}\sum_{j=0}^{T-1}(1-\mu\gamma)^{i+pj}\\
&=\sum_{i=0}^{Tp-1}(1-\mu\gamma)^i\\
&\leq\sum_{i=0}^{\infty}(1-\mu\gamma)^i\\
&\leq\frac{1}{\gamma\mu}.
\end{aligned}$$

Then,

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{x}_{K+1}^0-\boldsymbol{x}_*\|_2^2\\
&\leq(1-\gamma\mu)^{pK}\mathbb{E}\|\boldsymbol{x}_0^0-\boldsymbol{x}_*\|_2^2+\frac{1}{\gamma\mu}\left(L_{\max}\epsilon_{\mathrm{sfl}}^2\gamma+2\sigma^2\gamma^2\right)\\
&=(1-\gamma\mu)^{pK}\mathbb{E}\|\boldsymbol{x}_0^0-\boldsymbol{x}_*\|_2^2+\frac{\epsilon_{\mathrm{sfl}}^2}{\mu}+\frac{2\sigma^2}{\mu}\gamma.
\end{aligned}$$

Plugging in the value of $\epsilon_{\mathrm{sfl}}^2$ from Eq. (36), we have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{x}_{K+1}^0-\boldsymbol{x}_*\|_2^2\leq(1-\gamma\mu)^{pK}\mathbb{E}\|\boldsymbol{x}_0^0-\boldsymbol{x}_*\|_2^2\\
+\frac{L_{\max}n\sigma_{\mathrm{sub}}^2}{4b^2\mu}\gamma^2+\frac{2\sigma^2}{\mu}\gamma.
\end{aligned}$$

This implies

$$\mathbb{E}\|\boldsymbol{x}_{K+1}^0-\boldsymbol{x}_*\|_2^2\leq r^{Kn/b}\|\boldsymbol{x}_1^0-\boldsymbol{x}_*\|_2^2+C_{\mathrm{var}}^{\mathrm{sub}}\gamma^2+C_{\mathrm{var}}^{\mathrm{com}}\gamma,$$

where $r=1-\gamma\mu$,

$$C_{\mathrm{var}}^{\mathrm{sub}}=\frac{1}{4}\frac{L_{\max}}{\mu}\frac{n}{b^2}\left(\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\boldsymbol{x}_*)\|_2^2\right),\text{ and}$$

$$C_{\mathrm{var}}^{\mathrm{com}}=\frac{2}{\mu}\frac{n-b}{(n-1)b}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\right)+\frac{2}{\mu}\frac{n(b-1)}{(n-1)b}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i\right)^2.$$

Applying the fact that $(n-b)/n\leq(n-1)/n\leq 2$ for all $n\geq 2$ yields the simplified constants in the statement. □

B.5.3. Complexity Analysis (Theorem 4)

Theorem 4. *Let the objective F satisfy Assumption 1 and 2, where each component f_i is additionally μ-strongly convex, and Assumption 6 ($\mathrm{A}^{\mathrm{CVX}}$), 7 hold. Then, the last iterate $\boldsymbol{x}_T$ of doubly SGD-RR is ϵ-close to the global optimum $\boldsymbol{x}_* = \arg\max_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$ such that $\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}_*\|_2^2 \leq \epsilon$ after a number of iterations of at least*

$$T \geq \max\left(4C_{\mathrm{var}}^{\mathrm{com}}\frac{1}{\epsilon} + C_{\mathrm{var}}^{\mathrm{sub}}\frac{1}{\sqrt{\epsilon}},\, C_{\mathrm{bias}}\right)\log\left(2\left\|\boldsymbol{x}_1^0 - \boldsymbol{x}_*\right\|_2^2\frac{1}{\epsilon}\right)$$

for some fixed stepsize, where $T = Kp = Kn/b$,

$$C_{\mathrm{bias}} = \left(\mathcal{L}_{\max} + L\right)/\mu$$

$$C_{\mathrm{var}}^{\mathrm{com}} = \frac{2}{b}\left(\frac{1}{n}\sum_{i=1}^n\frac{\sigma_i^2}{\mu^2}\right) + 2\left(\frac{1}{n}\sum_{i=1}^n\frac{\sigma_i}{\mu}\right)^2,$$

$$C_{\mathrm{var}}^{\mathrm{sub}} = \sqrt{\frac{L_{\max}}{\mu}}\frac{\sqrt{n}}{b}\frac{\tau}{\mu}.$$

Proof. From the result of Theorem 5, we can invoke Lemma 5 with

$$A = \frac{L_{\max}n}{4b^2\mu}\tau^2,$$

$$B = \frac{2}{\mu}\left(\frac{n-b}{(n-1)b}\left(\frac{1}{n}\sum_{i=1}^n\sigma_i^2\right) + \frac{n(b-1)}{(n-1)b}\left(\frac{1}{n}\sum_{i=1}^n\sigma_i\right)^2\right),$$

$$C = \mathcal{L}_{\max} + L_{\max}.$$

Then, an ϵ accurate solution in expectation can be obtained after

$$T \geq \max\left(\underbrace{\frac{2B}{\mu}}_{\triangleq C_1}\frac{1}{\epsilon} + \underbrace{\frac{\sqrt{2A}}{\mu}}_{\triangleq C_2}\frac{1}{\sqrt{\epsilon}},\, \frac{\mathcal{L}_{\max} + L_{\max}}{\mu}\right)\log\left(2r_0^2\frac{1}{\epsilon}\right)$$

iterations with a stepsize of

$$\gamma = \min\left(\frac{-B + \sqrt{B^2 + 2A\epsilon}}{2A}, \frac{1}{C}\right).$$

To make the iteration complexity more precise, the terms C_1, C_2 can be organized as

$$C_1 = \frac{2B}{\mu} = \frac{2}{\mu}\left(\frac{2}{\mu}\left\{\frac{n-b}{(n-1)b}\left(\frac{1}{n}\sum_{i=1}^n\sigma_i^2\right) + \frac{n(b-1)}{(n-1)b}\left(\frac{1}{n}\sum_{i=1}^n\sigma_i\right)^2\right\}\right)$$

$$= \frac{4}{\mu^2}\left(\frac{n-b}{(n-1)b}\left(\frac{1}{n}\sum_{i=1}^n\sigma_i^2\right) + \frac{n(b-1)}{(n-1)b}\left(\frac{1}{n}\sum_{i=1}^n\sigma_i\right)^2\right)$$

$$C_2 = \frac{\sqrt{2A}}{\mu} = \sqrt{2\frac{L_{\max}n}{4b^2\mu}\tau^2}\frac{1}{\mu^2} = \frac{\sqrt{L_{\max}}\,\tau\sqrt{n}}{\sqrt{2}b\mu^{3/2}} \leq \frac{\sqrt{L_{\max}}}{\mu^{3/2}}\frac{\sqrt{n}}{b}\tau.$$

Applying the fact that $(n-b)/n \leq (n-1)/n \leq 2$ for all $n \geq 2$ yields the simplified constants in the statement. □

C. Applications

C.1. ERM with Randomized Smoothing

C.1.1. Description

Randomized smoothing was originally considered by Polyak & d Aleksandr Borisovich (1990); Nesterov (2005); Duchi et al. (2012) in the nonsmooth convex optimization context, where the function is "smoothed" through random perturbation. This scheme has recently renewed interest in the non-convex ERM context as it has been found to improve generalization performance (Orvieto et al., 2023; Liu et al., 2021). Here, we will focus on the computational aspect of this scheme. In particular, we will see if we can obtain similar computational guarantees already established in the finite-sum ERM setting, such as those by Gower et al. (2021a, Lemma 5.2).

Consider the canonical ERM problem, where we are given a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ and solve

$$\underset{\boldsymbol{w}\in\mathcal{W}}{\text{minimize}} \quad L(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(f_{\boldsymbol{w}}(\boldsymbol{x}_i), y_i\right) + h(\boldsymbol{w}),$$

where $(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ are the feature and label of the ith instance, $f_{\boldsymbol{w}} : \mathcal{X} \to \mathcal{Y}$ is the model, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is a non-negative loss function, and $h : \mathcal{W} \to \mathbb{R}$ is a regularizer.

For randomized smoothing, we instead minimize

$$L(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} R_i(\boldsymbol{w}),$$

where the instance risk is defined as

$$r_i(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{\epsilon}\sim\varphi}\ell\left(f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right)$$

for some noise distribution $\boldsymbol{\epsilon} \sim \varphi$. The goal is to obtain a solution $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}\in\mathcal{W}} L(\boldsymbol{w})$ that is robust to such perturbation.

The integrand of the gradient estimator of the instance risk is defined as

$$\begin{aligned}\boldsymbol{g}_i(\boldsymbol{w};\boldsymbol{\eta}) &= \nabla_{\boldsymbol{w}}\ell\left(f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right) \\ &= \frac{\partial f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i)}{\partial \boldsymbol{w}}\ell'\left(f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right),\end{aligned}$$

where it is an unbiased estimate of the instance risk such that

$$\mathbb{E}\boldsymbol{g}_i(\boldsymbol{w}) = \nabla R_i(\boldsymbol{w}).$$

The key challenge in analyzing the convergence of SGD in the ERM setting is dealing with the Jacobian $\frac{\partial f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i)}{\partial \boldsymbol{w}}$. Even for simple toy models, analyzing the Jacobian without relying on strong assumptions is hard. In this work, we will assume that it is bounded by an instance-dependent constant.

C.1.2. Preliminaries

We use the following assumptions:

Assumption 8.

(a) Let the mapping $\hat{y} \mapsto \ell(\hat{y}, y)$ is convex and L-smooth for any y_i $\forall i = 1, \ldots, n$.

(b) The Jacobian of the model with respect to its parameters for all $i = 1, \ldots, n$ is bounded almost surely as

$$\left\| \frac{f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i)}{\partial \boldsymbol{w}} \right\|_2 \leq G_i$$

for all $\boldsymbol{w} \in \mathcal{W}$.

(c) Interpolation holds on the solution set such that, for all $\boldsymbol{w}_* \in \arg\min_{\boldsymbol{w}\in\mathcal{W}} L(\boldsymbol{w})$, the loss minimized as

$$\ell\left(f_{\boldsymbol{w}_*+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right) = \ell'\left(f_{\boldsymbol{w}_*+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right) = 0$$

for all $(\boldsymbol{x}_i, y_i) \in \mathcal{D}$.

(a) holds for the squared loss, (c) basically assumes that the model is overparameterized and there exists a set of optimal weights that are robust with respect to perturbation. The has recently gained popularity as it qualitatively explains some of the empirical phenomenons of non-convex SGD (Vaswani et al., 2019; Gower et al., 2021a; Ma et al., 2018). (b) is a strong assumption but is commonly used to establish convergence guarantees of ERM (Gower et al., 2021a).

Remark 12. Under Assumption 8 (c), Assumption 7 holds with arbitrarily small σ_i^2, τ^2 .

C.1.3. Theoretical Analysis

Proposition 9. *Let Assumption 8 hold. Then, Assumption 6 (A^{ITP}) holds.*

Proof.

$$\mathbb{E}\|\boldsymbol{g}_i(\boldsymbol{w}) - \boldsymbol{g}_i(\boldsymbol{w}_*)\|_2^2$$
$$= \mathbb{E}\left\|\frac{\partial f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i)}{\partial \boldsymbol{w}}\ell'\left(f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right) - \frac{\partial f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i)}{\partial \boldsymbol{w}}\ell'\left(f_{\boldsymbol{w}_*+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right)\right\|_2^2,$$

from the interpolation assumption (Assumption 8 (c)),

$$= \mathbb{E}\left\|\frac{\partial f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i)}{\partial \boldsymbol{w}}\ell'\left(f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right)\right\|_2^2$$
$$\leq \mathbb{E}\left\|\frac{\partial f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i)}{\partial \boldsymbol{w}}\right\|_2^2\left|\ell'\left(f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right)\right|^2,$$

applying Assumption 8 (b),

$$\leq G_i^2\mathbb{E}\left|\ell'\left(f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right)\right|^2.$$

and then the interpolation assumption (Assumption 8 (c)),

$$= G_i^2\mathbb{E}\left|\ell'\left(f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right) - \ell'\left(f_{\boldsymbol{w}_*+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right)\right|.$$

From Assumption 8 (a),

$$\leq 2LG_i^2\mathbb{E}\Big(\ell\left(f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right) - \ell\left(f_{\boldsymbol{w}_*+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right) - \left\langle\ell'\left(f_{\boldsymbol{w}_*+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right), f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i) - f_{\boldsymbol{w}_*+\boldsymbol{\epsilon}}(\boldsymbol{x}_i)\right\rangle\Big)$$

and interpolation (Assumption 8 (c)),

$$= 2LG_i^2\left(\mathbb{E}\ell\left(f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right) - \mathbb{E}\ell\left(f_{\boldsymbol{w}_*+\boldsymbol{\epsilon}}(\boldsymbol{x}_i), y_i\right)\right)$$
$$= 2LG_i^2\left(R_i(\boldsymbol{w}) - R_i(\boldsymbol{w}_*)\right).$$

□

Proposition 10. *Let Assumption 8 hold. Then, Assumption 5 holds.*

Proof.

$$\frac{1}{n}\sum_{i=1}^n\|\nabla R_i(\boldsymbol{w}) - \nabla R_i(\boldsymbol{w}_*)\|_2^2$$
$$= \frac{1}{n}\sum_{i=1}^n\|\mathbb{E}\boldsymbol{g}_i(\boldsymbol{w}) - \mathbb{E}\boldsymbol{g}_i(\boldsymbol{w}_*)\|_2^2,$$

and from Jensen's inequality,

$$\leq \frac{1}{n}\sum_{i=1}^n\mathbb{E}\|\boldsymbol{g}_i(\boldsymbol{w}) - \boldsymbol{g}_i(\boldsymbol{w}_*)\|_2^2.$$

We can now reuse Proposition 9 as

$$\leq \frac{2}{n}\sum_{i=1}^n 2LG_i^2\left(R_i(\boldsymbol{w}) - R_i(\boldsymbol{w}_*)\right)$$

and taking $G_{\max} > G_i$ for all $i = 1, \ldots, n$ as

$$\leq 2LG_{\max}^2\frac{1}{n}\sum_{i=1}^n\left(R_i(\boldsymbol{w}) - R_i(\boldsymbol{w}_*)\right)$$
$$= 2LG_{\max}^2\left(L(\boldsymbol{w}) - L(\boldsymbol{w}_*)\right).$$

□

C.2. Reparameterization Gradient

C.2.1. Description

The reparameterization gradient estimator (Kingma & Welling, 2014; Rezende et al., 2014; Titsias & Lázaro-Gredilla, 2014) is a gradient estimator for problems of the form of

$$f_i(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{w}}} \ell_i(\boldsymbol{z}),$$

where $\ell : \mathbb{R}^{d_{\boldsymbol{z}}} \to \mathbb{R}$ is some integrand, such that the derivative is taken with respect to the parameters of the distribution$q_{\boldsymbol{w}}$ we are integrating over. It was independently proposed by Kingma & Welling (2014); Rezende et al. (2014) in the context of variational expectation maximization of deep latent variable models (a setup commonly known as variational autoencoders) and by Titsias & Lázaro-Gredilla (2014) for variational inference of Bayesian models.

Consider the case where the generative process of $q_{\boldsymbol{w}}$ can be represented as

$$\boldsymbol{z} \sim q_{\boldsymbol{w}} \quad \Leftrightarrow \quad \boldsymbol{z} \overset{d}{=} \mathcal{T}_{\boldsymbol{w}}(\boldsymbol{u}); \quad \boldsymbol{u} \sim \varphi,$$

where $\overset{d}{=}$ is equivalence in distribution, φ is some *base distribution* independent of $\boldsymbol{w}$, and $\mathcal{T}_{\boldsymbol{w}}$ is a *reparameterization function* measurable with respect to φ and differentiable with respect to all $\boldsymbol{w} \in \mathcal{W}$. Then, the reparameterization gradient is given by the integrand

$$\boldsymbol{g}_i(\boldsymbol{w}; \boldsymbol{u}) = \nabla_{\boldsymbol{w}} \ell_i(\mathcal{T}_{\boldsymbol{w}}(\boldsymbol{u})),$$

which is unbiased, and often results in lower variance (Kucukelbir et al., 2017; Xu et al., 2019) compared to alternatives such as the score gradient estimator. (See Mohamed et al. (2020) for an overview of such estimators.)

The reparameterization gradient is primarily used to solve problems in the form of

$$\begin{aligned}\underset{\boldsymbol{w}\in\mathcal{W}}{\text{minimize}} \quad F(\boldsymbol{w}) &= \sum_{i=1}^{n} f_i(\boldsymbol{w}) + h(\boldsymbol{w}) \\ &= \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{w}}} \ell_i(\boldsymbol{z}) + h(\boldsymbol{w}),\end{aligned}$$

where h is some convex regularization term.

Previously, Domke (2019, Theorem 6) established a bound on the gradient variance of the reparameterization gradient (Kingma & Welling, 2014; Rezende et al., 2014; Titsias & Lázaro-Gredilla, 2014) under the doubly stochastic setting. This bound also incorporates more advanced subsampling strategies such as importance sampling Gower et al. (2019); Gorbunov et al. (2020); Csiba & Richtárik (2018); Needell et al. (2016); Needell & Ward (2017). However, he did not extend the analysis to a complexity analysis of SGD and left out the effect of correlation between components.

C.2.2. Preliminaries

The properties of the reparameterization gradient for when $q_{\boldsymbol{w}}$ is in the location-scale family were studied by Domke (2019).

Assumption 9. We assume the variational family

$$\mathcal{Q} \triangleq \{q_{\boldsymbol{w}} \mid \boldsymbol{w} \in \mathcal{W}\}$$

satisfies the following:

(a) $\mathcal{Q}$ is part of the location-scale family such that $\mathcal{T}_{\boldsymbol{w}}(\boldsymbol{u}) = \boldsymbol{C}\boldsymbol{u} + \boldsymbol{m}$.

(b) The scale matrix is positive definite such that $\boldsymbol{C} \succ 0$.

(c) $\boldsymbol{u} = (u_1, \ldots, u_{d_{\boldsymbol{z}}})$ constitute of *i.i.d.* components, where each component is standardized, symmetric, and finite kurtosis such that $\mathbb{E}u_i = 0$, $\mathbb{E}u_i^2 = 1$, $\mathbb{E}u_i^3 = 0$, and $\mathbb{E}u_i^4 = k_\varphi$, where k_φ is the kurtosis.

Under these conditions, Domke (2019) proves the following:

Lemma 13 (Domke, 2019; Theorem 3)**.** *Let Assumption 9 hold and ℓ_i be L_i-smooth. Then, the squared norm of the reparameterization gradient is bounded:*

$$\mathbb{E}\|\boldsymbol{g}_i(\boldsymbol{w})\|_2^2 \le (d+1)\|\boldsymbol{m} - \bar{\boldsymbol{z}}_i\|_2^2 + \left(d + k_\varphi\right)\|\boldsymbol{C}\|_{\mathrm{F}}^2$$

for all $\boldsymbol{w} = (\boldsymbol{m}, \boldsymbol{C}) \in \mathcal{W}$ and all stationary points of ℓ_i denoted with $\bar{\boldsymbol{z}}_i$.

Similarly, Kim et al. (2023) establish the QES condition as part of Lemma 3 (Kim et al., 2023). We refine this into statement we need:

Lemma 14. *Let Assumption 9 hold and ℓ_i be L_i-smooth. Then, the squared norm of the reparameterization gradient is bounded:*

$$\mathbb{E}\left\|\boldsymbol{g}_i(\boldsymbol{w}) - \boldsymbol{g}_i(\boldsymbol{w}')\right\|_2^2 \leq L_i^2\left(d + k_\varphi\right)\|\boldsymbol{w} - \bar{\boldsymbol{w}}_i\|_2^2$$

for all $\boldsymbol{w}, \boldsymbol{w}' \in \mathcal{W}$.

Proof.

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{g}_i(\boldsymbol{w}) - \boldsymbol{g}_i(\boldsymbol{w}')\right\| \\
&= \mathbb{E}\|\nabla_{\boldsymbol{w}}\ell_i(\mathcal{T}_{\boldsymbol{w}}(\boldsymbol{u})) - \nabla_{\boldsymbol{w}}\ell_i(\mathcal{T}_{\boldsymbol{w}'}(\boldsymbol{u}))\|_2^2 \\
&= \mathbb{E}\left\|\frac{\partial\mathcal{T}_{\boldsymbol{w}}(\boldsymbol{u})}{\partial\boldsymbol{w}}\nabla\ell_i(\mathcal{T}_{\boldsymbol{w}}(\boldsymbol{u})) - \frac{\partial\mathcal{T}_{\boldsymbol{w}'}(\boldsymbol{u})}{\partial\boldsymbol{w}'}\nabla\ell_i(\mathcal{T}_{\boldsymbol{w}'}(\boldsymbol{z}))\right\|_2^2 \\
&= \mathbb{E}(\nabla\ell_i(\mathcal{T}_{\boldsymbol{w}}(\boldsymbol{u})) - \nabla\ell_i(\mathcal{T}_{\boldsymbol{w}'}(\boldsymbol{u})))^\top\left(\frac{\partial\mathcal{T}_{\boldsymbol{w}}(\boldsymbol{u})}{\partial\boldsymbol{w}}\right)^\top\frac{\partial\mathcal{T}_{\boldsymbol{w}'}(\boldsymbol{u})}{\partial\boldsymbol{w}'} \\
&\quad\times(\nabla\ell_i(\mathcal{T}_{\boldsymbol{w}}(\boldsymbol{u})) - \nabla\ell_i(\mathcal{T}_{\boldsymbol{w}'}(\boldsymbol{u}))).
\end{aligned}$$

As shown by Kim et al. (2023, Lemma 6), the squared Jacobian is an identity matrix scaled with a scalar-valued function independent of $\boldsymbol{w}$, $J_{\mathcal{T}}(\boldsymbol{u}) = \|\boldsymbol{u}\|_2^2 + 1$, such that

$$= \mathbb{E}J_{\mathcal{T}}(\boldsymbol{u})\|\nabla\ell_i(\mathcal{T}_{\boldsymbol{w}}(\boldsymbol{u})) - \nabla\ell_i(\mathcal{T}_{\boldsymbol{w}'}(\boldsymbol{u}))\|_2^2,$$

applying the L_i-smoothness of ℓ_i,

$$= L_i^2\mathbb{E}J_{\mathcal{T}}(\boldsymbol{u})\|\mathcal{T}_{\boldsymbol{w}}(\boldsymbol{u}) - \mathcal{T}_{\boldsymbol{w}'}(\boldsymbol{u})\|_2^2,$$

and Kim et al. (2023, Corollary 2) show that,

$$\leq L_i^2\left(d + k_\varphi\right)\|\boldsymbol{w} - \boldsymbol{w}'\|_2^2.$$

□

Lastly, the properties of ℓ_i are known to transfer to the expectation f_i as follows:

Lemma 15. *Let Assumption 9 hold. Then we have the following:*

(i) Let ℓ_i be L_i smooth. Then, f_i is also L_i-smooth

(ii) Let ℓ_i be convex. Then, f_i and F are also convex.

(iii) Let ℓ_i be μ-strongly convex. Then, f_i and F are also μ-strongly convex.

Proof. (i) is proven by Domke (2020, Theorem 1), while a more general result is provided by Kim et al. (2023, Theorem 1); (ii) and (iii) are proven by Domke (2020, Theorem 9) and follow from the fact that h is convex. □

C.2.3. Theoretical Analysis

We now conclude that the reparameterization gradient fits the framework of this work:

Proposition 11. *Let Assumption 9 hold and ℓ_i be convex and L_i-smooth. Then, Assumption 5 holds.*

Proof. The result follows from combining Lemma 15 and Lemma 10. □

From Lemma 13, we satisfy 7.

Proposition 12. *Let Assumption 9 hold, ℓ_i be L_i-smooth, the solutions $\boldsymbol{w}_* \in \arg\min_{\boldsymbol{w}\in\mathcal{W}} F(\boldsymbol{w})$ and the stationary points of ℓ_i, $\bar{\boldsymbol{z}}$, be bounded such that $\|\boldsymbol{w}_*\|_2 < \infty$ and $\|\bar{\boldsymbol{z}}\|_2 < \infty$. Then, Assumption 7 holds.*

Proof. Lemma 13 implies that, as long as the $\boldsymbol{w}_*$ and $\bar{\boldsymbol{z}}$ are bounded, we satisfy the component gradient estimator part of Assumption 7, where the constant is given as

$$\sigma_i^2 = L_i^2(d+1)\|\boldsymbol{m}_* - \bar{\boldsymbol{z}}_i\|_2^2 + L_i^2\left(d + k_\varphi\right)\|\boldsymbol{C}_*\|_{\mathrm{F}}^2,$$

where $\boldsymbol{w}_* = (\boldsymbol{m}_*, \boldsymbol{C}_*)$. □

From Lemma 14, we can conclude that the reparameterization gradient satisfies Assumption 6:

Proposition 13. *Let Assumption 9 hold and ℓ_i be L_i-smooth and μ-strongly convex. Then, Assumption 6 ($\mathrm{A}^{\mathrm{CVX}}$) and Assumption 6 (B) hold.*

Proof. Notice the following:

1. Assumption 4 always holds for $\rho = 1$.
2. From the stated conditions, Lemma 15 establishes that both f_i and F are μ-strongly convex.
3. μ-strong convexity of f and F implies that both are μ-QFG (Karimi et al., 2016, Appendix A).
4. The reparameterization gradient satisfies the QES condition by Lemma 14.

Item 1, 2 and 3 combined imply the ES condition by Proposition 8, which immediately implies the ER condition with the same constant. Therefore, we satisfy both Assumption 6 ($\mathrm{A}^{\mathrm{CVX}}$), Assumption 6 (B) where the ER constant $\mathcal{L}_i$ is given as

$$\mathcal{L}_i = \frac{L_i^2}{\mu}\left(d + k_\varphi\right).$$

□