On Biased Stochastic Gradient Estimation

Derek Driggs*, Jingwei Liang†, and Carola-Bibiane Schönlieb‡

Department of Applied Mathematics and Theoretical Physics, Cambridge University

February 28, 2020

Abstract

We present a uniform analysis of biased stochastic gradient methods for minimizing convex, strongly convex, and non-convex composite objectives, and identify settings where bias is useful in stochastic gradient estimation. The framework we present allows us to extend proximal support to biased algorithms, including SAG and SARAH, for the first time in the convex setting. We also use our framework to develop a new algorithm, Stochastic Average Recursive GradiEnt (SARGE), that achieves the oracle complexity lower-bound for non-convex, finite-sum objectives and requires strictly fewer calls to a stochastic gradient oracle per iteration than SVRG and SARAH. We support our theoretical results with numerical experiments that demonstrate the benefits of certain biased gradient estimators.

1 Introduction

In this paper, we focus on the following composite minimisation problem:

$$\min_{x\in\mathbb{R}^p} \big\{F(x) \stackrel{\text{def}}{=} f(x) + g(x)\big\}. \tag{1}$$

Throughout, we assume:

- g is proper and closed such that its proximity operator (see (3) in Section 2) is well posed,
- f admits finite-sum structure $f(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n f_i(x)$, and for all $i \in \{1, 2, \cdots, n\}$, ∇f_i is L-Lipschitz continuous for some $L > 0$.

We place no further restrictions on f_i or g unless stated otherwise.

Problems of this form arise frequently in many areas of science and engineering, such as machine learning, statistics, operations research, and imaging. For instance, in machine learning, these problems often arise as empirical risk minimisation problems from classification and regression tasks. Examples include ridge regression, logistic regression, Lasso, and ℓ_1-regularized logistic regression [8]. Principal component analysis (PCA) can also be formulated as a problem with this structure, where the functions f_i are non-convex [6, 16]. In imaging, ℓ_1 or total variation regularization is often combined with differentiable data discrepancy terms that appear in both convex and non-convex instances [10].

1.1 Stochastic gradient methods

Two classical approaches to solve (1) are the proximal gradient descent method (PGD) [21] and its accelerated variants, including inertial PGD [20] and FISTA [7]. For these deterministic approaches, the full gradient of f must be evaluated at each iteration, which often requires huge computational resources when n is large. Such a drawback makes these schemes unsuitable for large-scale machine learning tasks.

*d.driggs@damtp.cam.ac.uk
†jl993@cam.ac.uk
‡cbs31@cam.ac.uk

By exploiting the finite sum structure of f, stochastic gradient methods enjoy low per-iteration complexity while achieving comparable convergence rates. These qualities make stochastic gradient methods the standard approach to solving many problems in machine learning, and are gaining popularity in other areas such as image processing [11]. Stochastic gradient descent (SGD) was first proposed in the 1950's [28] and has experienced a renaissance in the past decade, with numerous variants of SGD proposed in the literature (see, for instance, [13, 18, 30] and references therein). Most of these algorithms can be summarised into one general form, which is described below in Algorithm 1.

Algorithm 1 Stochastic gradient descent framework

Input: starting point $x_0 \in \mathbb{R}^p$, gradient estimator $\widetilde{\nabla}$.

1: **for** $k = 0, 1, \cdots, T-1$ **do**

2: Compute stochastic gradient $\widetilde{\nabla}_k$ at current iteration k.

3: Choose step size/learning rate η_k.

4: Update x_{k+1}:

$$x_{k+1} \leftarrow \text{prox}_{\eta_k g}(x_k - \eta_k \widetilde{\nabla}_k). \tag{2}$$

5: **end for**

Below we summarize several popular stochastic gradient estimators $\widetilde{\nabla}_k$:

- **SGD** Classic stochastic gradient descent [28] uses the following gradient estimator at iteration k:

$$\begin{cases} \text{sample } j_k \text{ uniformly at random from } \{1, ..., n\}, \\ \widetilde{\nabla}_k^{\text{SGD}} = \nabla f_{j_k}(x_k). \end{cases}$$

  At each step, SGD uses the gradient of the sampled function $\nabla f_{j_k}(x_k)$ as a stochastic approximation of the full gradient $\nabla f(x_k)$. It is an unbiased estimate as $\mathbb{E}_k[\nabla f_{j_k}(x_k)] = \nabla f(x_k)$. It is also *memoryless*: every update of x_{k+1} depends only upon x_k and the random variable j_k. The variance of SGD is does not vanish as x_k converges.

- **SAG** To deal with the non-vanishing variance of SGD, in [29, 30] the authors introduce the SAG gradient estimator, which uses the gradient history to decrease its variance. With $\nabla f_i(\varphi_0^i) = 0, i = 1, ..., n$, the SAG gradient estimator is computed using the following procedure:

$$\begin{cases} \text{sample } j_k \text{ uniformly at random from } \{1, ..., n\}, \\ \widetilde{\nabla}_k^{\text{SAG}} = \frac{1}{n}(\nabla f_{j_k}(x_k) - \nabla f_{j_k}(\varphi_k^{j_k})) + \frac{1}{n}\sum_{i=1}^n \nabla f_i(\varphi_k^i), \\ \text{update the gradient history: } \varphi_{k+1}^{j_k} = x_k \text{ and } \nabla f_i(\varphi_{k+1}^i) = \begin{cases} \nabla f_i(x_k) & \text{if } \ i = j_k, \\ \nabla f_i(\varphi_k^i) & \text{o.w.} \end{cases} \end{cases}$$

  Here, for each $i \in \{1, ..., n\}$, $\nabla f_i(\varphi_k^i)$ is a stored gradient of ∇f_i from a previous iteration. With the help of memory, the variance of the SAG gradient estimator diminishes as x_k approaches the solution; estimators that satisfy this property are known as *variance-reduced* estimators. In contrast to the SGD estimator, $\widetilde{\nabla}_k^{\text{SAG}}$ is a *biased* estimate of $\nabla f(x_k)$.

- **SAGA** Based on [29, 30], [13] propose the unbiased gradient estimator SAGA, which is computed using the procedure below.

$$\begin{cases} \text{Sample } j_k \text{ uniformly at random from } \{1, ..., n\}, \\ \widetilde{\nabla}_k^{\text{SAGA}} = \nabla f_{j_k}(x_k) - \nabla f_{j_k}(\varphi_k^{j_k}) + \frac{1}{n}\sum_{i=1}^n \nabla f_i(\varphi_k^i), \\ \text{update the gradient history: } \varphi_{k+1}^{j_k} = x_k \text{ and } \nabla f_i(\varphi_{k+1}^i) = \begin{cases} \nabla f_i(x_k) & \text{if } \ i = j_k, \\ \nabla f_i(\varphi_k^i) & \text{o.w.} \end{cases} \end{cases}$$

  Compared to $\widetilde{\nabla}^{\text{SAG}}$, the SAGA estimator gives less weight to stored gradients. With this adjustment, $\widetilde{\nabla}^{\text{SAGA}}$ is unbiased while maintaining the variance reduction property. Similar gradient estimators can be found in Point-SAGA [12], Finito [14], MISO [22], SDCA [31], and those in [17].

- **SVRG** Another popular variance-reduced estimator is SVRG [18]. The SVRG gradient estimator is computed as follows:

$$\begin{bmatrix} \text{For } s = 0, \cdots, S \\ \nabla f(\varphi_s) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\varphi_s), \\ \text{For } k = 1, \cdots, m \\ \quad \begin{bmatrix} \text{Sample } j_k \text{ uniformly at random from } \{1, \cdots, n\}, \\ \widetilde{\nabla}_k^{\text{SVRG}} = \nabla f_{j_k}(x_k) - \nabla f_{j_k}(\varphi_s) + \nabla f(\varphi_s), \end{bmatrix} \end{bmatrix}$$

  where φ_s is a "snapshot" point updated every m steps. The algorithms prox-SVRG [34], Katyusha [2], KatyushaX [3], Natasha [1], Natasha2 [4], MiG [35], ASVRG [32], and VARAG [19] use the SVRG gradient estimator.

- **SARAH** In [24] the authors proposed a recursive modification to SVRG.

$$\begin{bmatrix} \text{For } s = 0, \cdots, S \\ \widetilde{\nabla}_{k-1}^{\text{SARAH}} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\varphi_s), \\ \text{For } k = 1, \cdots, m \\ \quad \begin{bmatrix} \text{Sample } j_k \text{ uniformly at random from } \{1, \cdots, n\}, \\ \widetilde{\nabla}_k^{\text{SARAH}} = \nabla f_{j_k}(x_k) - \nabla f_{j_k}(x_{k-1}) + \widetilde{\nabla}_{k-1}^{\text{SARAH}}, \end{bmatrix} \end{bmatrix}$$

  Like SAG, SARAH is a biased gradient estimator. It is also used in prox-SARAH [25], SPIDER [15], SPIDERBoost [33] and SPIDER-M [36].

We refer to algorithms employing (un)biased gradient estimators as (un)biased stochastic algorithms, respectively. The body of work on biased algorithms is stunted compared to the enormous literature on unbiased algorithms, leading to several gaps in the development of biased stochastic gradient methods. We list a few below.

- **Complex convergence proofs.** It is commonly believed that the relationship $\mathbb{E}_k[\widetilde{\nabla}_k] = \nabla f(x_k)$ is essential for a simple convergence analysis (see, for example, the discussion in [13]). The convergence proof of the biased algorithm SAG is especially complex, requiring computational verification [29, 30].
- **Sub-optimal convergence rates.** In the convex setting with $g \equiv 0$, SARAH achieves a complexity bound of $\mathcal{O}(\frac{\log(1/\epsilon)}{\epsilon})$ [24] for finding a point $\bar{x}_k$ such that $\mathbb{E}[F(\bar{x}_k) - F(x^\star)] \leq \epsilon$. In comparison, SAGA and SVRG achieve a complexity bound of $\mathcal{O}(1/\epsilon)$ which is the same as deterministic proximal gradient descent.
- **Lack of proximal support.** Bias also makes it difficult to handle non-smooth functions. To the best of our knowledge, there are no theoretical guarantees for biased algorithms to solve (1) with $g \not\equiv 0$ that take advantage of convexity when it is present.

Despite the above shortcomings, there are notable exceptions that suggest biased algorithms are worth further consideration. Recently, [15, 25, 33, 36] proved that algorithms using the SARAH gradient estimator require $\mathcal{O}(\sqrt{n}/\epsilon^2)$ stochastic gradient evaluations to find an ϵ-first order stationary point. This matches the complexity lower-bound for non-convex, finite-sum optimisation for smooth functions f_i and $n \leq \mathcal{O}(\epsilon^{-4})$ [15]. For comparison, the best complexity bound obtained for SAGA and SVRG in this setting is $\mathcal{O}(n^{2/3}/\epsilon^2)$ [5, 26]. A detailed summary of existing complexity bounds for the variance-reduced gradient estimators mentioned above is provided in Table 1 for convex, strongly convex and non-convex settings.

1.2 Contributions

This work provides three main contributions:

1. We introduce a framework for the systematic analysis of a large class of stochastic gradient methods and investigate a bias-variance tradeoff arising from our analysis. Our analysis provides proximal support to biased algorithms for the first time in the convex setting.

	Convex	Strongly Convex	Non-Convex	Proximal Support?
SAGA	$\mathcal{O}(\frac{nL}{\epsilon})$	$\mathcal{O}((n+\kappa)\log(1/\epsilon))$	$\mathcal{O}(\frac{nL}{\epsilon^2})^b$	Yes
SVRG	$\mathcal{O}(\frac{nL}{\epsilon})^a$	$\mathcal{O}((n+\kappa)\log(1/\epsilon))$	$\mathcal{O}(\frac{nL}{\epsilon^2})^b$	Yes
SAG	$\mathcal{O}(\frac{nL}{\epsilon})$	$\mathcal{O}((n+\kappa)\log(1/\epsilon))$	**Unknown**	**No**
SARAH	$\mathcal{O}(\frac{L\log(1/\epsilon)}{\epsilon})$	$\mathcal{O}((n+\kappa)\log(1/\epsilon))$	$\mathcal{O}(\frac{\sqrt{n}L}{\epsilon^2})$	**Non-Convex Only**

[a]The algorithm SVRG++ reduces this rate to $\mathcal{O}(n\log(1/\epsilon)+L/\epsilon)$ using an epoch-doubling procedure [6].
[b]Mini-batching reduces the dependence on n to $n^{2/3}$ [5, 26], and these rates are proven only in the case g is convex.

Table 1: Existing complexity bounds for stochastic gradient estimators under different settings. These complexities represent the number of stochastic gradient oracle calls required to find a point satisfying $\mathbb{E}[F(\bar{x}_k) - F(x^\star)] \leq \epsilon$ for the convex case, $\mathbb{E}[\|x_k - x^\star\|^2] \leq \epsilon$ for the strongly convex case, and an ϵ-approximate stationary point (as in Definition 1) in the non-convex case. In the strongly convex case, μ is the strong convexity constant, and $\kappa = L/\mu$ is the condition number.

2. We apply our framework to derive convergence rates for SARAH and biased versions of SAGA and SVRG on convex, strongly convex, and non-convex problems.
3. We design a new recursive gradient estimator, Stochastic Average Recursive GradiEnt (SARGE), that achieves the same convergence rates as SARAH but never computes a full gradient, giving it a strictly smaller per-iteration complexity than SARAH. In particular, we show that SARGE achieves the oracle complexity lower bound for non-convex finite-sum optimisation.

To study the effects of bias on the SAGA and SVRG estimators, we introduce Biased SAGA (B-SAGA) and Biased SVRG (B-SVRG). For $\theta > 0$ the B-SAGA and B-SVRG gradient estimators are

- B-SAGA:
$$\widetilde{\nabla}_k^{\text{B-SAGA}} \overset{\text{def}}{=} \tfrac{1}{\theta}(\nabla f_{j_k}(x_k) - \nabla f_{j_k}(\varphi_k^{j_k})) + \tfrac{1}{n}\sum_{i=1}^n \nabla f_i(\varphi_k^i),$$
- B-SVRG:
$$\widetilde{\nabla}_k^{\text{B-SVRG}} \overset{\text{def}}{=} \tfrac{1}{\theta}(\nabla f_{j_k}(x_k) - \nabla f_{j_k}(\varphi_s)) + \nabla f(\varphi_s).$$

In both B-SAGA and B-SVRG, the bias parameter θ adjusts how much weight is given to stored gradient information. When $\theta = n$, $\widetilde{\nabla}_k^{\text{B-SAGA}}$ recovers the SAG gradient estimator.

Motivated by the desirable properties of SARAH, we propose a new gradient estimator, Stochastic Average Recursive GradiEnt (SARGE), which is defined below

$$\widetilde{\nabla}_k^{\text{SARGE}} \overset{\text{def}}{=} \nabla f_{j_k}(x_k) - \psi_k^{j_k} + \tfrac{1}{n}\sum_{i=1}^n \psi_k^i - (1-\tfrac{1}{n})(\nabla f_{j_k}(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}),$$

where the variables ψ_k^i follow the update rule $\psi_{k+1}^{i_k} = \nabla f_{j_k}(x_k) - (1-\frac{1}{n})\nabla f_{j_k}(x_{k-1})$. Similar to SAGA, SARGE uses stored gradient information to avoid having to compute the full gradient, a computational burden that SVRG and SARAH require for variance reduction.

A summary of the complexity results obtained from our analysis for SAG/B-SAGA, B-SVRG, SARAH, and SARGE are provided in Table 2. Note that the result for SAG is included in B-SAGA.

Paper organization Preliminary results and notations are provided in Section 2. A discussion on the tradeoff between bias and variance in stochastic optimisation is included in Section 3. Our main convergence results are presented in Section 4. In Section 5, we substantiate our theoretical results using numerical experiments involving several classic regression tasks arising from machine learning. All the proofs of the main results are collected in the appendix.

	Convex	Strongly Convex	Non-Convex	Proximal Support?
B-SAGA[c]	$\mathcal{O}(\frac{nL}{\epsilon})$	$\mathcal{O}(n\kappa\log(1/\epsilon))$	$\mathcal{O}(\frac{nL}{\epsilon^2})$	Yes
B-SVRG[c]	$\mathcal{O}(\frac{nL}{\epsilon})$	$\mathcal{O}(n\kappa\log(1/\epsilon))$	$\mathcal{O}(\frac{nL}{\epsilon^2})$	Yes
SARAH	$\mathcal{O}(\frac{\sqrt{n}L}{\epsilon})$	$\mathcal{O}(\max\{\sqrt{n}\kappa, n\}\log(1/\epsilon))$	$\mathcal{O}(\frac{\sqrt{n}L}{\epsilon^2})$	Yes
SARGE	$\mathcal{O}(\frac{\sqrt{n}L}{\epsilon})$	$\mathcal{O}(\max\{\sqrt{n}\kappa, n\}\log(1/\epsilon))$	$\mathcal{O}(\frac{\sqrt{n}L}{\epsilon^2})$	Yes

[c]Mini-batching reduces the dependence on n to $n^{2/3}$ as in [5, 26] giving these algorithms a lower complexity than full-gradient methods, but we do not consider mini-batching in this work.

Table 2: Complexity bounds obtained from our analysis framework. These complexities represent the number of stochastic gradient oracle calls required to find a point $\bar{x}_k$ satisfying $\mathbb{E}[F(\bar{x}_k) - F(x^\star)] \leq \epsilon$ for the convex case, $\mathbb{E}[\|x_k - x^\star\|^2] \leq \epsilon$ for the strongly convex case, and an ϵ-approximate stationary point in the non-convex case.

2 Preliminaries and notations

Throughout the paper, $\mathbb{R}^p$ is a p-dimensional Euclidean space equipped with scalar inner product $\langle \cdot, \cdot \rangle$ and associated norm $\|\cdot\|$. The sub-differential of a proper closed convex function g is the set-valued operator defined by $\partial g \stackrel{\text{def}}{=} \{v \in \mathbb{R}^n | g(x') \geq g(x) + \langle v, x' - x\rangle, \forall x' \in \mathbb{R}^n\}$, the proximal map of g is defined as

$$\text{prox}_{\eta g}(y) \stackrel{\text{def}}{=} \arg\min_{x \in \mathbb{R}^n} \eta g(x) + \tfrac{1}{2}\|x - y\|^2, \tag{3}$$

where $\eta > 0$ and $y \in \mathbb{R}^p$. With $y^+ = \text{prox}_{\gamma g}(y)$, (3) is equivalent to $y - y^+ \in \eta \partial g(y^+)$.

Below we summarize some useful results in convex analysis.

Lemma 1 ([23, Thm 2.1.5]) *Suppose f is convex with an L-Lipschitz continuous gradient. We have for every $x, u \in \mathbb{R}^p$,*

$$\|\nabla f(x) - \nabla f(u)\|^2 \leq 2L(f(x) - f(u) - \langle \nabla f(u), x - u\rangle).$$

When f is a finite sum as in (1), Lemma 1 is equivalent to the following result.

Lemma 2 *Let $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$, where each f_i is convex with an L-Lipschitz gradient. Then for every $x, u \in \mathbb{R}^p$,*

$$\tfrac{1}{2Ln}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f_i(u)\|^2 \leq f(x) - f(u) - \langle \nabla f(u), x - u\rangle.$$

Lemma 2 is obtained by applying Lemma 1 to each f_i and averaging.

Lemma 3 *Suppose g is μ-strongly convex with $\mu \geq 0$, and let $z = \text{prox}_{\eta g}(x - \eta d)$ for some $x, d \in \mathbb{R}^p$ and $\eta > 0$. Then, for any $y \in \mathbb{R}^p$,*

$$\eta\langle d, z - y\rangle \leq \tfrac{1}{2}\|x - y\|^2 - \tfrac{1+\mu\eta}{2}\|z - y\|^2 - \tfrac{1}{2}\|z - x\|^2 - \eta g(z) + \eta g(y).$$

Proof. By the strong convexity of g,

$$g(z) - g(y) \leq \langle \xi, z - y\rangle - \tfrac{\mu}{2}\|z - y\|^2, \quad \forall \xi \in \partial g(z).$$

From the definition of the proximal operator, we have that $\frac{1}{\eta}(x - z) - d \in \partial g(z)$. Therefore,

$$\begin{aligned} g(z) - g(y) &\leq \langle \xi, z - y\rangle - \tfrac{\mu}{2}\|z - y\|^2 \\ &= \tfrac{1}{\eta}\langle x - z - \eta d, z - y\rangle - \tfrac{\mu}{2}\|z - y\|^2 \\ &= -\langle d, z - y\rangle + \tfrac{1}{\eta}\langle x - z, z - y\rangle - \tfrac{\mu}{2}\|z - y\|^2 \\ &= -\langle d, z - y\rangle - \tfrac{1}{2\eta}\|x - z\|^2 - \tfrac{1}{2\eta}\|z - y\|^2 + \tfrac{1}{2\eta}\|x - y\|^2 - \tfrac{\mu}{2}\|z - y\|^2. \end{aligned}$$

Multiplying by η and rearranging yields the assertion. □

The next lemma is an analogue of the descent lemma for gradient descent when the gradient is replaced with an arbitrary vector d.

Lemma 4 *Suppose g is μ-strongly convex for $\mu \geq 0$, and let $z = \text{prox}_{\eta g}(x - \eta d)$. The following inequality holds for any $\lambda > 0$.*

$$0 \leq \eta(F(x) - F(z)) + \tfrac{\eta}{2L\lambda}\|d - \nabla f(x)\|^2 + (\tfrac{\eta L(\lambda+1)}{2} - \tfrac{2+\mu\eta}{2})\|z - x\|^2.$$

Proof. This follows immediately from Lemma 3.

$$\begin{aligned}
0 &= \eta\langle d, x - z\rangle + \eta\langle d, z - x\rangle \\
&\overset{①}{\leq} \eta\langle d, x - z\rangle - \tfrac{2+\mu\eta}{2}\|z - x\|^2 + \eta(g(x) - g(z)) \\
&= \eta\langle \nabla f(x), x - z\rangle + \eta\langle d - \nabla f(x), x - z\rangle - \tfrac{2+\mu\eta}{2}\|z - x\|^2 + \eta(g(x) - g(z)) \\
&\overset{②}{\leq} \eta(F(x) - F(z)) + \eta\langle d - \nabla f(x), x - z\rangle + (\tfrac{\eta L}{2} - \tfrac{2+\mu\eta}{2})\|z - x\|^2 \\
&\overset{③}{\leq} \eta(F(x) - F(z)) + \tfrac{\eta}{2L\lambda}\|d - \nabla f(x)\|^2 + (\tfrac{\eta L(\lambda+1)}{2} - \tfrac{2+\mu\eta}{2})\|z - x\|^2.
\end{aligned}$$

Inequality ① is due to Lemma 3 with $y = x$, ② is due to the Lipschitz continuity of ∇f_i, and ③ is Young's. □

The previous two lemmas require g to be convex. Similar results hold in the non-convex case as well.

Lemma 5 *Let $z = \text{prox}_{\eta g}(x - \eta d)$ for some $x, d \in \mathbb{R}^p$ and $\eta > 0$. Then, for any $y \in \mathbb{R}^p$,*

$$\eta\langle d, z - y\rangle \leq \tfrac{1}{2}\|x - y\|^2 - \tfrac{1}{2}\|z - x\|^2 - \eta g(z) + \eta g(y).$$

Proof. By the Lipschitz continuity of ∇f, we have the inequalities

$$\begin{aligned}
f(x) - f(y) &\leq \langle \nabla f(x), x - y\rangle + \tfrac{L}{2}\|x - y\|^2, \\
f(z) - f(x) &\leq \langle \nabla f(x), z - x\rangle + \tfrac{L}{2}\|z - x\|^2.
\end{aligned}$$

Furthermore, by the definition of z,

$$z \in \arg\min_v \left\{ \langle d, v - x\rangle + \frac{1}{2\eta}\|v - x\|^2 + g(v) \right\}.$$

Taking $v = y$, we obtain

$$g(z) - g(y) \leq \langle d, y - z\rangle + \tfrac{1}{2\eta}\left(\|x - y\|^2 - \|x - z\|^2\right).$$

Adding these three inequalities and multiplying by η completes the proof. □

Lemma 6 *Let $z = \text{prox}_{\eta g}(x - \eta d)$. Then*

$$F(z) \leq F(y) + \langle \nabla f(x) - d, z - y\rangle + (\tfrac{L}{2} - \tfrac{1}{2\eta})\|x - z\|^2 + (\tfrac{L}{2} + \tfrac{1}{2\eta})\|x - y\|^2.$$

Proof. By the Lipschitz continuity of ∇f, we have the inequalities

$$\begin{aligned}
f(x) - f(y) &\leq \langle \nabla f(x), x - y\rangle + \tfrac{L}{2}\|x - y\|^2, \\
f(z) - f(x) &\leq \langle \nabla f(x), z - x\rangle + \tfrac{L}{2}\|z - x\|^2.
\end{aligned}$$

Furthermore, by Lemma 5,

$$g(z) - g(y) \leq \langle d, y - z\rangle + \tfrac{1}{2\eta}\left(\|x - y\|^2 - \|x - z\|^2\right).$$

Adding these inequalities together completes the proof. □

In the non-convex setting, to measure convergence of the sequence to a first-order stationary point, we use the notion of a generalized gradient [23].

Definition 1 (Generalized gradient map) *The generalized gradient map is defined as*

$$\mathcal{G}_{\eta_k}(x_k) \stackrel{\text{def}}{=} \tfrac{1}{\eta_k}(x_k - \text{prox}_{\eta_k g}(x_k - \eta_k \nabla f(x_k))).$$

When $g \equiv 0$, we have $\mathcal{G}_{\eta_k}(x_k) = \nabla f(x_k) \to 0$ if the sequence $\{x_k\}$ converges to some $x^\star \in \mathbb{R}^p$ such that $\nabla f(x^\star) = 0$. For nontrivial g, suppose $\inf_k \eta_k > 0$ and x_k converges to some $x^\star$ such that $x^\star = \text{prox}_{\eta g}(x^\star - \eta \nabla f(x^\star))$, then $\mathcal{G}_{\eta_k}(x) \to 0$. Such a point $x^\star$ is called *first-order stationary point* of (1) and an *ϵ-first-order stationary point* is a point satisfying $\|\mathcal{G}_\eta(x)\| \le \epsilon$ for some $\eta > 0$.

3 A bias-variance tradeoff in stochastic gradient methods

In this section, we discuss the effect of the bias and variance of a stochastic gradient estimator on the performance of Algorithm 1. It is elementary that the mean-squared error (MSE) of a stochastic estimator can be decomposed into the sum of its variance and squared bias. In our setting,

$$\mathbb{E}_k[\|\widetilde{\nabla}_k - \nabla f(x_k)\|^2] = \|\mathbb{E}_k[\widetilde{\nabla}_k] - \nabla f(x_k)\|^2 + \mathbb{E}_k[\|\widetilde{\nabla}_k - \mathbb{E}_k[\widetilde{\nabla}_k]\|^2].$$

This decomposition shows that a biased estimator might have a smaller MSE than an unbiased estimator as long as the bias sufficiently diminishes the variance. This is the *bias-variance tradeoff*. As we see below, a bias-variance tradeoff exists in our analysis of stochastic gradient methods, but with a slightly different form.

In what follows, we first discuss the bias-variance tradeoff in the convex settings and then the non-convex setting.

3.1 Convex case

Let $x^\star$ be a global minimizer of problem (1). From the update (2), let $w_{k+1} \in \partial g(x_{k+1})$. We have the following bound on the suboptimality at x_{k+1}:

$$\begin{aligned}
&\mathbb{E}_k[F(x_{k+1}) - F(x^\star)] \\
&= \mathbb{E}_k[f(x_{k+1}) - f(x_k) + f(x_k) - f(x^\star) + g(x_{k+1}) - g(x^\star)] \\
&\overset{①}{\le} \tfrac{L}{2}\mathbb{E}_k[\|x_{k+1} - x_k\|^2] + \mathbb{E}_k[\langle \nabla f(x_k), x_{k+1} - x_k\rangle] + \langle \nabla f(x_k), x_k - x^\star\rangle + \mathbb{E}_k[g(x_{k+1}) - g(x^\star)] \\
&= \tfrac{L}{2}\mathbb{E}_k[\|x_{k+1} - x_k\|^2] + \mathbb{E}_k[\langle \nabla f(x_k) - \widetilde{\nabla}_k, x_{k+1} - x_k\rangle] \\
&\quad + \mathbb{E}_k[\langle \nabla f(x_k) - \widetilde{\nabla}_k, x_k - x^\star\rangle] + \mathbb{E}_k[\langle \widetilde{\nabla}_k, x_{k+1} - x^\star\rangle] + \mathbb{E}_k[g(x_{k+1}) - g(x^\star)] \\
&\overset{②}{\le} \tfrac{\epsilon}{2}\mathbb{E}_k[\|\nabla f(x_k) - \widetilde{\nabla}_k\|^2] + (\tfrac{L}{2} + \tfrac{1}{2\epsilon})\mathbb{E}_k[\|x_{k+1} - x_k\|^2] \\
&\quad + \mathbb{E}_k[\langle \nabla f(x_k) - \widetilde{\nabla}_k, x_k - x^\star\rangle] + \mathbb{E}_k[\langle \widetilde{\nabla}_k + w_{k+1}, x_{k+1} - x^\star\rangle - \tfrac{\mu}{2}\|x_{k+1} - x^*\|^2] \\
&\overset{③}{\le} \tfrac{\epsilon}{2}\mathbb{E}_k[\|\nabla f(x_k) - \widetilde{\nabla}_k\|^2] + (\tfrac{L}{2} + \tfrac{1}{2\epsilon} - \tfrac{1}{2\eta})\mathbb{E}_k[\|x_{k+1} - x_k\|^2] \\
&\quad + \mathbb{E}_k[\langle \nabla f(x_k) - \widetilde{\nabla}_k, x_k - x^\star\rangle] - \tfrac{1+\mu\eta}{2\eta}\mathbb{E}_k[\|x_{k+1} - x^\star\|^2] + \tfrac{1}{2\eta}\|x_k - x^\star\|^2.
\end{aligned} \tag{4}$$

Inequality ① follows from the convexity of f and Lipschitz continuity of ∇f, ② follows from the (strong) convexity of g, and ③ comes from the implicit definition of the proximal operator (3). For the last line of the inequality, we observe that the inner product term $\mathbb{E}_k[\langle \nabla f(x_k) - \widetilde{\nabla}_k, x_k - x^\star\rangle]$ vanishes when $\widetilde{\nabla}_k$ is an unbiased estimate of $\nabla f(x_k)$. When the estimator is biased, we must develop a new way to control this term, together with $\mathbb{E}_k[\|\nabla f(x_k) - \widetilde{\nabla}_k\|^2]$.

Hence, the following terms arise in our convergence analysis from the bias and the variance of the gradient estimator:

$$\begin{aligned}
\text{Bias} &: \mathbb{E}_k[\langle \nabla f(x_k) - \widetilde{\nabla}_k, x_k - x^\star\rangle] \text{ and } \|\mathbb{E}_k[\widetilde{\nabla}_k] - \nabla f(x_k)\|^2, \\
\text{Variance} &: \mathbb{E}_k[\|\widetilde{\nabla}_k - \mathbb{E}_k[\widetilde{\nabla}_k]\|^2].
\end{aligned}$$

Remark 1 (Non-composite case $g = 0$) *When $g = 0$, for gradient descent, the descent property of f yields*

$$f(x_{k+1}) - f(x^\star) \leq (\tfrac{L}{2} - \tfrac{1}{\eta})\|x_{k+1} - x_k\|^2 + f(x_k) - f(x^\star),$$

where $\eta \leq 2/L$. For stochastic gradient descent, we obtain the following relationship:

$$\begin{aligned}
&\mathbb{E}_k[f(x_{k+1}) - f(x^\star)] \\
&= \mathbb{E}_k[f(x_{k+1}) - f(x_k) + f(x_k) - f(x^\star)] \\
&\leq \mathbb{E}_k[\langle \nabla f(x_k) - \widetilde{\nabla}_k, x_{k+1} - x_k\rangle] + (\tfrac{L}{2} - \tfrac{1}{\eta})\mathbb{E}_k[\|x_{k+1} - x_k\|^2] + f(x_k) - f(x^\star) \\
&\leq \tfrac{\epsilon}{2}\mathbb{E}_k[\|\nabla f(x_k) - \widetilde{\nabla}_k\|^2] + (\tfrac{L}{2} + \tfrac{1}{2\epsilon} - \tfrac{1}{\eta})\mathbb{E}_k[\|x_{k+1} - x_k\|^2] + f(x_k) - f(x^\star).
\end{aligned} \tag{5}$$

Compared to (4)*, there is no inner product term in* (5)*, which makes the analysis of the non-composite case much simpler. This is one reason why biased algorithms have been successfully studied in non-composite setting, but not in the composite setting.*

3.2 Non-convex case

The influence of bias is simpler in the non-convex setting and independent of g, which explains why biased algorithms have recently found success for these problems. To begin, let $\hat{x}_{k+1} \overset{\text{def}}{=} \operatorname{prox}_{\eta/2g}(x_k - \eta/2\nabla f(x_k))$. Applying Lemma 6 with $z = \hat{x}_{k+1}$, $y = x = x_k$ and $d = \nabla f(x_k)$, we have

$$F(\hat{x}_{k+1}) \leq F(x_k) + (\tfrac{L}{2} - \tfrac{1}{\eta})\|\hat{x}_{k+1} - x_k\|^2.$$

Again, applying Lemma 6 with $z = x_{k+1}$, $y = \hat{x}_{k+1}$, $x = x_k$, and $d = \widetilde{\nabla}_k$, we obtain

$$\begin{aligned}
F(x_{k+1}) \leq F(\hat{x}_{k+1}) &+ \langle \nabla f(x_k) - \widetilde{\nabla}_k, x_{k+1} - \hat{x}_{k+1}\rangle + (\tfrac{L}{2} - \tfrac{1}{2\eta})\|x_{k+1} - x_k\|^2 \\
&+ (\tfrac{L}{2} + \tfrac{1}{2\eta})\|\hat{x}_{k+1} - x_k\|^2
\end{aligned}$$

Adding these two inequalities together gives

$$\begin{aligned}
F(x_{k+1}) &\leq F(x_k) + (L - \tfrac{1}{2\eta})\|\hat{x}_{k+1} - x_k\|^2 + (\tfrac{L}{2} - \tfrac{1}{2\eta})\|x_{k+1} - x_k\|^2 \\
&\quad + \langle \nabla f(x_k) - \widetilde{\nabla}_k, x_{k+1} - \hat{x}_{k+1}\rangle \\
&\overset{①}{\leq} F(x_k) + (L - \tfrac{1}{2\eta})\|\hat{x}_{k+1} - x_k\|^2 + (\tfrac{L}{2} - \tfrac{1}{2\eta})\|x_{k+1} - x_k\|^2 + 2\eta\|\nabla f(x_k) - \widetilde{\nabla}_k\|^2 \\
&\quad + \frac{1}{8\eta}\|\hat{x}_{k+1} - x_{k+1}\|^2 \\
&\overset{②}{\leq} F(x_k) + (L - \tfrac{1}{4\eta})\|\hat{x}_{k+1} - x_k\|^2 + (\tfrac{L}{2} - \tfrac{1}{4\eta})\|x_{k+1} - x_k\|^2 + 2\eta\|\nabla f(x_k) - \widetilde{\nabla}_k\|^2.
\end{aligned} \tag{6}$$

Inequality ① is Young's, and ② is the standard inequality $\|a-c\|^2 \leq 2\|a-b\|^2 + 2\|b-c\|^2$. In the non-convex case, the inner-product bias term does not appear, so the bias-variance tradeoff is the classical one.

3.3 General bounds on bias and variance

To ensure convergence for a particular gradient estimator used in Algorithm 1, we must bound the inner-product bias term $\mathbb{E}_k[\langle \nabla f(x_k) - \widetilde{\nabla}_k, x_k - x^\star\rangle]$ and the MSE $\mathbb{E}_k[\|\nabla f(x_k) - \widetilde{\nabla}_k\|^2]$. Below we introduce general bounds on these terms that allow us to establish convergence rates for a variety of gradient estimators. The first of these is a bound on the MSE term.

Definition 2 (Bounded MSE) *The stochastic gradient estimator $\widetilde{\nabla}$ is said to satisfy the* BMSE$(M_1, M_2, \rho_M, \rho_F, m)$ *property with parameters $M_1, M_2 \geq 0$, $\rho_M, \rho_F \in (0, 1]$ and $m \geq 1$ if there exist sequences $\mathcal{M}_k$ and $\mathcal{F}_k$ such that*

$$\sum_{k=ms}^{m(s+1)-1} \mathbb{E}[\|\widetilde{\nabla}_k - \nabla f(x_k)\|^2] \leq \mathcal{M}_{ms},$$

and the following bounds hold:

$$\mathcal{M}_{ms} \leq (1-\rho_M)^m \mathcal{M}_{m(s-1)} + \mathcal{F}_{ms} + \tfrac{M_1}{n} \sum_{k=ms}^{m(s+1)-1} \sum_{i=1}^{n} \mathbb{E}[\|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\|^2],$$

$$\mathcal{F}_{ms} \leq \sum_{\ell=0}^{s} \tfrac{M_2(1-\rho_F)^{m(s-\ell)}}{n} \sum_{k=ms}^{m(s+1)-1} \sum_{i=1}^{n} \mathbb{E}[\|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\|^2].$$

The constant m is the epoch length of the gradient estimator, hence it is usually set to be $\mathcal{O}(n)$. The BMSE property allows these bounds to hold only on average over an epoch. This property is useful in convergence analyses because it bounds the MSE by a geometrically decaying sequence $\{\mathcal{M}_{mk}\}_{k\in\mathbb{N}}$ and a component that is proportional to the one-iteration progress of gradient descent ($1/n \sum_{i=1}^n \|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\|^2$).

Remark 2

- *Most variance-reduced stochastic gradient estimators satisfy the BMSE property, including SAG, SAGA, SVRG, SARAH, and all the estimators in [17]. SGD does not satisfy this property, as its variance does not decay along the iterations.*
- *Most existing work on the analysis of general stochastic gradient algorithms enforce bounds of this form on either the MSE or the moments of the stochastic estimator, with the crucial difference that existing works require the bounds to (i.e., dependent on only the previous iteration) [9]. In contrast, the BMSE property allows non-Markovian MSE bounds through the sequence $\mathcal{F}_k$. This relaxation is crucial for the analysis of our new gradient estimator, SARGE.*

In order to bound the inner-product bias term, we require the gradient estimator to admit a certain structure in its bias. In biased estimators such as SAG, the bias depends on the stored gradient values:

$$\nabla f(x_k) - \mathbb{E}_k[\widetilde{\nabla}_k^{\text{SAG}}] = (1-\tfrac{1}{n})\big(\nabla f(x_k) - \tfrac{1}{n}\textstyle\sum_{i=1}^n \nabla f_i(\varphi_k^i)\big).$$

We call estimators whose bias admits the above structure *memory-biased* gradient estimators. These include SAG, and more generally B-SAGA and B-SVRG.

Definition 3 (Memory-biased gradient estimator) *The stochastic gradient estimator $\widetilde{\nabla}$ is* memory-biased *with parameters $\theta > 0$, $B_1 \geq 0$, and $m \geq 1$ if*

$$\nabla f(x_k) - \mathbb{E}_k[\widetilde{\nabla}_k] = (1-\tfrac{1}{\theta})\Big(\nabla f(x_k) - \frac{1}{n}\sum_{i=1}^n \nabla f_i(\varphi_k^i)\Big),$$

for some $\{\varphi_k^i\}_{i=1}^n \subset \{x_\ell\}_{\ell=0}^{k-1}$, and for any $s \in \mathbb{N}_0$,

$$\sum_{k=ms}^{m(s+1)-1} \tfrac{1}{n}\sum_{i=1}^n \mathbb{E}[\|x_k - \varphi_k^i\|^2] \leq B_1 \sum_{k=ms}^{m(s+1)-1} \mathbb{E}[\|x_k - x_{k-1}\|^2]. \tag{7}$$

B-SAGA is clearly a memory-biased estimator, and so is B-SVRG where $\varphi_k^i = \varphi_{ms}^i$ for all k in epoch s. The parameter θ controls the amount of bias in the estimator, and B_1, in a sense, measures how "stale" the stored gradient information is. For memory-biased gradient estimators, the bias-term can be handled easily.

Lemma 7 *Suppose $\widetilde{\nabla}$ is memory-biased with parameter $\theta \geq 1$ and that F is μ-strongly convex with $\mu \geq 0$. For any $\lambda > 0$, the following inequality holds:*

$$\begin{aligned}\eta\mathbb{E}_k[F(x_{k+1}) - F(x^\star)] \leq \tfrac{\eta}{2L\lambda}\mathbb{E}_k[\|\widetilde{\nabla}_k - \nabla f(x_k)\|^2] - \tfrac{1+\mu\eta}{2}\mathbb{E}_k[\|x_{k+1} - x^\star\|^2] + \tfrac{1}{2}\|x_k - x^\star\|^2 \\ + (\tfrac{\eta L(\lambda+1)}{2} - \tfrac{1}{2})\mathbb{E}_k[\|x_{k+1} - x_k\|^2] + \tfrac{\eta L}{2n}(1-\tfrac{1}{\theta})\sum_{i=1}^n \|x_k - \varphi_k^i\|^2.\end{aligned}$$

The proof of Lemma 7 can be found in Appendix A. The bound of Lemma 7 is analogous to the bound in (4), but the inner-product bias term is replaced with $\frac{\eta L}{2n}(1-\frac{1}{\theta})\sum_{i=1}^{n}\|x_k-\varphi_k^i\|^2$. This term is proportional to the progress of gradient descent (by (7)), so this provides the necessary control over the inner-product bias term.

For estimators such as SARAH, the bias depends on the error in the previous gradient estimate, rather than previous stochastic gradients:

$$\nabla f(x_k)-\mathbb{E}_k[\widetilde{\nabla}_k^{\text{SARAH}}]=\nabla f(x_{k-1})-\widetilde{\nabla}_{k-1}^{\text{SARAH}}.$$

We refer to estimators of this type as *recursively biased*.

Definition 4 (Recursively biased gradient estimator) *For any sequence $\{x_k\}$, let $\widetilde{\nabla}_k$ be a stochastic gradient estimator generated from the points $\{x_\ell\}_{\ell=0}^k$. This estimator is* recursively biased *with parameters $\rho_B\in(0,1]$ and $\nu\geq 1$ if*

$$\nabla f(x_k)-\mathbb{E}_k[\widetilde{\nabla}_k]=\begin{cases}0 & \text{for } k\in\nu\mathbb{N}_0,\\ (1-\rho_B)(\nabla f(x_{k-1})-\widetilde{\nabla}_{k-1}) & \text{o.w.}\end{cases}$$

The parameter ν represents how many steps occur between full gradient evaluations. For SARGE, $\nu=\infty$ because the full gradient is never computed.

Lemma 8 *Suppose $\widetilde{\nabla}$ is a recursively biased gradient estimator with parameters $\nu\geq 1$ and $\rho_B\in(0,1]$. Then, for any $\epsilon>0$,*

$$\begin{aligned}&\sum_{k=\nu s+1}^{\nu(s+1)-1}|\mathbb{E}\langle\nabla f(x_{k-1})-\widetilde{\nabla}_{k-1},x_k-x^\star\rangle|\\ &\leq\min\left\{\nu,\tfrac{1}{\rho_B}\right\}\sum_{k=\nu s}^{\nu(s+1)-1}\mathbb{E}\left[\tfrac{\epsilon}{2}\|\nabla f(x_{k+1})-\widetilde{\nabla}_{k+1}\|^2+\tfrac{1}{2\epsilon}\|x_{k+1}-x_k\|^2\right].\end{aligned}$$

Lemma 8 shows that, for recursively biased estimators, the inner-product bias term $\langle\nabla f(x_{k-1})-\widetilde{\nabla}_{k-1},x_k-x^\star\rangle$ is bounded from above by the MSE, implying that introducing bias to decrease the MSE is a reasonable approach to design improved gradient estimators.

4 Convergence rates

In this section, we analyse the convergence rates for the stochastic gradient methods. We first provide very general convergence rates based on the bounds from the last section. Then, we specify the result to specific gradient estimators including memory-biased B-SAGA/B-SVRG, and recursively biased SARAH and SARGE.

4.1 General convergence rates

For Algorithm 1, we consider a constant step size $\eta_k\equiv\eta>0$. Given T iterations of Algorithm 1, define the average iterate $\bar{x}_T\overset{\text{def}}{=}1/T\sum_{k=1}^{T}x_k$.

4.1.1 Convex and strongly convex cases

The following theorem establishes convergence rates for memory-biased estimators in the convex regime.

Theorem 9 (Memory-biased estimators) *Let $\widetilde{\nabla}$ be a memory-biased gradient estimator parameterized by $\theta\geq 1$ and $B_1\geq 0$, which satisfies the BMSE$(M_1,M_2,\rho_M,\rho_F,m)$ property. Let $\Theta=\frac{M_1\rho_F+2M_2}{\rho_M\rho_F}$ and $\rho=\min\{\rho_M,\rho_F\}$.*

- *When F is convex, let $\eta = \frac{1}{L(1+3\sqrt{2\Theta})}$, then*

$$\mathbb{E}[F(\bar{x}_T) - F(x^\star)] \leq \frac{1}{T}\left(\frac{L(1+3\sqrt{2\Theta})\|x_0 - x^\star\|^2}{2} + \max\left\{\frac{B_1(1-1/\theta)}{\sqrt{2\Theta}} - 1, 0\right\}\frac{F(x_0)-F(x^\star)}{L(1+3\sqrt{2\Theta})}\right).$$

- *When F is μ-strongly convex with $\mu > 0$, let $\eta = \min\left\{\frac{1}{3L(1+3\sqrt{2\Theta})}, \frac{\sqrt{2\Theta}}{B_1\mu(1-1/\theta)}, \frac{\rho}{2\mu}\right\}$. The iterate x_T satisfies*

$$\mathbb{E}[\|x_T - x^\star\|^2] \leq (1+\mu\eta)^{-T}\left(\tfrac{2}{\mu}(F(x_0) - F(x^\star)) + \|x_0 - x^\star\|^2\right).$$

The proof of Theorem 9 is provided in Appendix A. The next result establishes convergence rates for recursively biased gradient estimators whose proof is in Appendix B.

Theorem 10 (Recursively biased estimators) *Let $\widetilde{\nabla}$ be a recursively biased gradient estimator parameterized by $\rho_B \in (0,1)$ and $\nu \geq 1$, which satisfies the BMSE$(M_1, M_2, \rho_M, \rho_F, m)$ property. Let $B_2 \stackrel{\text{def}}{=} \min\{\nu, 1/\rho_B\}$, $\Theta = \frac{M_1\rho_F + 2M_2}{\rho_M\rho_F}$ and $\rho = \min\{\rho_M, \rho_F\}$.*

- *When F is convex, let $\eta = \frac{1}{L(4\sqrt{2\Theta}+1)}$, then*

$$\mathbb{E}[F(\bar{x}_T) - F(x^\star)] \leq \frac{1}{T}\left(\frac{L(4\sqrt{2\Theta}+1)}{2}\|x_0 - x^\star\|^2 + \max\left\{(1-\rho_B)B_2 - 1, 0\right\}\frac{F(x_0)-F(x^\star)}{L(4\sqrt{2\Theta}+1)}\right).$$

- *When F is μ-strongly convex with $\mu > 0$, let $\eta = \min\left\{\frac{1}{3L(4\sqrt{2\Theta}+1)}, \frac{1}{\mu(1-\rho_B)B_2}, \frac{\rho}{2\mu}\right\}$, then*

$$\mathbb{E}[\|x_T - x^\star\|^2] \leq (1+\mu\eta)^{-T}\left(\tfrac{2}{\mu}(F(x_0) - F(x^\star)) + \|x_0 - x^\star\|^2\right).$$

Remark 3

- *Both theorems hold true for smaller η; the choices in the theorems are the largest ones allowed by our analysis.*
- *For B-SAGA and B-SVRG, $\Theta = \mathcal{O}(n^2)$, while for SARAH and SARGE, $\Theta = \mathcal{O}(n)$. This gives these recursive gradient estimators improved convergence rates and suggests that the bias in these estimators is more effective than the bias in SAGA and SVRG.*

4.1.2 Non-convex case

The analysis of biased gradient estimators is simpler for the non-convex setting than the convex ones due to the absence of the inner-product bias term in (6). Below we provide a uniform convergence guarantee for all gradient estimators satisfying the BMSE property, regardless of their bias. This suggests that in the non-convex setting, a large-bias, small-MSE gradient estimator is favourable over an estimator with small bias and large MSE.

Theorem 11 *Let $\widetilde{\nabla}$ be a gradient estimator that satisfies the BMSE$(M_1, M_2, \rho_M, \rho_F, m)$ property, let $\Theta = \frac{M_1\rho_F + 2M_2}{\rho_M\rho_F}$, and let α be a chosen uniformly at random from the set $\{0, 1, \cdots, T-1\}$. If F is non-convex, set $\eta = \frac{\sqrt{16\Theta+1}-1}{16L\Theta}$ in Algorithm 1, and the point x_α satisfies the following bound on its generalized gradient:*

$$\mathbb{E}[\|\mathcal{G}_{\eta/2}(x_\alpha)\|^2] \leq \frac{16(F(x_0)-F(x^\star))}{T\eta(1-4\eta L)}.$$

The proof of this result is provided in Appendix C.

Remark 4 *The convergence result of Theorem 11 does not depend on the bias except through the MSE of the gradient estimator, which implies that incorporating arbitrary amounts of bias for a smaller MSE improves the convergence rate. This fact is what allows the recursively biased estimators SARAH and SARGE to achieve the oracle complexity lower bound for non-convex optimisation when they are used in Algorithm 1.*

4.2 Convergence rates for specific gradient estimators

In this section, we specialise the general convergence rates to analyse the performance of B-SAGA, B-SVRG, SARAH, and SARGE.

4.2.1 Biased SAGA and SVRG

B-SAGA and B-SVRG are examples of memory-biased gradient estimators, as their biases take the form

$$\nabla f(x_k) - \mathbb{E}_k[\widetilde{\nabla}_k] = (1 - \tfrac{1}{\theta})\Big(\nabla f(x_k) - \tfrac{1}{n}\sum_{i=1}^{n} \nabla f_i(\varphi_k^i)\Big),$$

for some previous iterates φ_k^i. To establish convergence rates for B-SAGA and B-SVRG, we only need to show these estimators satisfy the BMSE property with suitable constants.

Lemma 12 *The B-SAGA gradient estimator is memory-biased with $B_1 = 2n(2n+1)$, and it satisfies the BMSE property with parameters $\rho_M = \frac{1}{2n}$, $m = 1$, $M_2 = 0$, $\rho_F = 1$, and*

$$M_1 = \begin{cases} \frac{2n+1}{\theta^2} & \theta \in (0, 2], \\ (2n+1)(1-\frac{1}{\theta})^2 & \theta > 2. \end{cases}$$

The proof of Lemma 12 uses a slight modification of existing variance bounds for the SAGA estimator, appearing in [13], for example. We include the proof in Appendix D. The B-SVRG gradient estimator satisfies the BMSE property with similar constants.

Lemma 13 *The B-SVRG gradient estimator is memory-biased with $B_1 = 3m(m+1)$, and it satisfies the BMSE property with parameters $\rho_M = 1$, $M_2 = 0$, $\rho_F = 1$, and*

$$M_1 = \begin{cases} \frac{3m(m+1)}{\theta^2} & \theta \in (0, 2], \\ 3m(m+1)(1-\frac{1}{\theta})^2 & \theta > 2. \end{cases}$$

With these constants established, Theorem 9 provides rates of convergence.[1]

Corollary 14 (Convergence rates for B-SAGA) *Algorithm 1 achieves the following convergence guarantees using the B-SAGA gradient estimator:*

- *If F is convex, depending on the choice of θ, set the step size to*

$$\eta = \eta_\theta \stackrel{\text{def}}{=} \begin{cases} \frac{1}{L(1+\frac{6}{\theta}\sqrt{n(2n+1)})} : \theta \in [1, 2], \\ \frac{1}{L(1+6(1-\frac{1}{\theta})\sqrt{n(2n+1)})} : \theta > 2, \end{cases}$$

  and $\bar{x}_T$ satisfies $\mathbb{E}[F(\bar{x}_T) - F(x^\star)] = \mathcal{O}(Ln/T)$.
- *If F is μ-strongly convex, set $\eta = \min\left\{\eta_\theta, \frac{1}{4\mu n}\right\}$. Then x_T satisfies $\mathbb{E}[\|x_T - x^\star\|^2] = \mathcal{O}((1+\mu\eta)^{-T})$.*
- *If F is non-convex, after T iterations, the generalized gradient at x_α satisfies*

$$\mathbb{E}[\|\mathcal{G}_{\eta/2}(x_\alpha)\|^2] = \begin{cases} \mathcal{O}\left(\frac{Ln}{T\theta}\right) : \eta = \frac{\theta}{2L\sqrt{n(2n+1)}}, \ \theta \in (0, 2], \\ \mathcal{O}\left(\frac{Ln}{T(1-\frac{1}{\theta})}\right) : \eta = \frac{1}{2L(1-\frac{1}{\theta})\sqrt{n(2n+1)}}, \ \theta > 2. \end{cases}$$

Corollary 15 (Convergence rates of B-SVRG) *Algorithm 1 achieves the following convergence guarantees using the B-SVRG gradient estimator:*

- *When F is convex, depending on the choice of θ, set the step size to*

$$\eta = \eta_\theta = \begin{cases} \frac{1}{L(1+\frac{3}{\theta}\sqrt{6m(m+1)})} : \theta \in [1, 2], \\ \frac{1}{L(1+3(1-\frac{1}{\theta})\sqrt{6m(m+1)})} : \theta > 2. \end{cases}$$

  After S epochs, the point $\bar{x}_{mS}$ satisfies $\mathbb{E}[F(\bar{x}_{mS}) - F(x^\star)] = \mathcal{O}(L/S)$.

[1] We state the convergence rates without constants for simplicity. The complete result with constants is included in Appendix D.

- *If, moreover, F is μ-strongly convex, let $\eta = \min\{\eta_\theta, \frac{1}{2\mu}\}$. After S epochs, x_{mS} satisfies $\mathbb{E}[\|x_{mS} - x^\star\|^2] = \mathcal{O}((1+\mu\eta)^{-mS})$.*
- *If F is non-convex, after S epochs, the generalized gradient at x_α satisfies*

$$\mathbb{E}[\|\mathcal{G}_{\eta/2}(x_\alpha)\|^2] = \begin{cases} \mathcal{O}\left(\frac{Lm}{T\theta}\right) : \eta = \frac{\sqrt{2}\theta}{2L\sqrt{3m(m+1)}}, \ \theta \in (0,2], \\ \mathcal{O}\left(\frac{Lm}{T(1-1/\theta)}\right) : \eta = \frac{\sqrt{2}\theta}{2L(1-\frac{1}{\theta})\sqrt{3m(m+1)}}, \ \theta > 2. \end{cases}$$

Remark 5

- *Our MSE bounds and convergence rates are optimised when $\theta = 2$. Numerical experiments (including those in Section 5) suggest that setting θ in the range $1 < \theta \ll n$ gives the best performance, and B-SAGA prefers larger values of θ than B-SVRG.*
- *In the special case $\theta = 1$, Corollaries 14 and 15 recover the state-of-the-art rates for SAGA and SVRG in the convex and non-convex regimes. For strongly convex problems, these rates are worse than existing convergence rates of $\mathcal{O}((1+\min\left\{\frac{\mu}{L}, \frac{1}{n}\right\})^{-T})$ proven for SAGA and SVRG [13, 34]. This difference is due to the generality of Theorem 9, as some memory-biased estimators, including B-SVRG, exhibit poor performance on strongly convex problems when the bias is large.*
- *Corollaries 14 and 15 require step sizes that decrease with n, while existing results for SAG, SAGA, and SVRG allow step sizes that are independent of n. This is also due to the generality of Theorem 9. For example, we find in practice that B-SAGA converges with step sizes that are independent of n, but B-SVRG requires smaller step sizes when the epoch length is larger.*

4.2.2 SARAH and SARGE

The SARAH and SARGE gradient estimators are recursively biased, with

$$\nabla f(x_k) - \mathbb{E}_k[\widetilde{\nabla}_k^{\text{SARAH}}] = \nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARAH}}.$$

and

$$\nabla f(x_k) - \mathbb{E}_k[\widetilde{\nabla}_k^{\text{SARGE}}] = (1-\tfrac{1}{n})(\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}).$$

As we shall see, these biased estimators admit smaller MSE bounds than unbiased and memory-biased estimators, and this is reflected in their improved convergence rates. The following two lemmas establish the constants appearing in Theorem 10 for these estimators.

Lemma 16 *The SARAH gradient estimator is recursively biased with parameters $\rho_B = 0$ and $\nu = m$, and it satisfies the BMSE property with parameters $M_1 = m$, $\rho_M = 1$, $\rho_F = 1$, and $M_2 = 0$.*

Lemma 17 *The SARGE gradient estimator is recursively biased with parameters $\rho_B = 1/n$ and $\nu = \infty$, and it satisfies the BMSE property with $M_1 = 12$, $M_2 = 39/n$, $\rho_M = \frac{1}{4n}$, $\rho_F = \frac{1}{2n}$, and $m = 1$.*

Proofs of these results are included in Appendices E and F, respectively. It is enlightening to compare these BMSE constants to those of B-SVRG and B-SAGA. M_1 is a factor of n smaller for the SARAH and SARGE estimators than for the B-SVRG and B-SAGA estimators (as long as $m = \mathcal{O}(n)$ in SARAH and B-SVRG). This translates to an $\mathcal{O}(\sqrt{n})$ improvement in the convergence rates for SARAH and SARGE.

Corollary 18 (Convergence rates for SARAH) *When using the SARAH gradient estimator in Algorithm 1,*

- *If F is convex, set $\eta = \frac{1}{L(4\sqrt{2m}+1)}$. After T iterations, $\bar{x}_T$ satisfies $\mathbb{E}[F(\bar{x}_T) - F(x^\star)] = \mathcal{O}(L\sqrt{m}/T)$.*
- *If F is μ-strongly convex, set $\eta = \min\{\frac{1}{3L(4\sqrt{2m}+1)}, \frac{1}{\mu m}\}$, then $\mathbb{E}[\|x_T - x^\star\|^2] = \mathcal{O}((1+\mu\eta)^{-T})$.*
- *If F is non-convex, set $\eta = \frac{1}{L\sqrt{2m}}$, then $\mathbb{E}[\|\mathcal{G}_{\eta/2}(x_\alpha)\|^2] \leq \mathcal{O}\left(L\sqrt{m}/T\right)$*

Corollary 19 (Convergence rates for SARGE) *When using the SARGE gradient estimator in Algorithm 1,*

- *If F is convex, set $\eta = \frac{1}{L(16\sqrt{3(n+13)}+1)}$, then $\mathbb{E}[F(\bar{x}_T) - F(x^\star)] = \mathcal{O}(L\sqrt{n}/T)$.*

- *If F is μ-strongly convex, set* $\eta = \min\{\frac{1}{3L(16\sqrt{3(n+13)}+1)}, \frac{1}{4\mu n}\}$, *then* $\mathbb{E}[\|x_T - x^\star\|^2] = \mathcal{O}((1+\mu\eta)^{-T})$.
- *If F is non-convex, set* $\eta = \frac{1}{4L\sqrt{3(n+13)}}$, *then* $\mathbb{E}[\|\mathcal{G}(x_\alpha)\|^2] \leq \mathcal{O}\left(L\sqrt{n}/T\right)$.

These convergence rates for convex objectives represent a significant improvement over the performance of SAGA, SVRG, and full-gradient methods. Each of these algorithms require $\mathcal{O}(\frac{nL}{\epsilon})$ stochastic gradient evaluations to find a point satisfying $F(x_T) - F(x^\star) \leq \epsilon$, while SARAH and SARGE require only $\mathcal{O}(\frac{\sqrt{n}L}{\epsilon})$. These rates do not require the epoch-doubling procedure of [6], although epoch-doubling can potentially be used to improve the performance of SARAH just as it improves the performance of SVRG on non-strongly convex objectives.

This square-root dependence on n is present in the convergence rates for strongly convex and non-convex objectives as well, which is a significant improvement over the dependence on n in the convergence rates of B-SAGA and B-SVRG. This better dependence on n is most significant in the non-convex regime, where these convergence rates imply that the SARAH and SARGE gradient estimators require only $\mathcal{O}(\frac{\sqrt{n}L}{\epsilon})$ stochastic gradient evaluations to find an ϵ-approximate stationary point, which is the oracle-complexity lower bound [15]. Similar results already exist for algorithms using the SARAH estimator [15, 25, 33, 36]. Our results for SARGE show that achieving this complexity is possible without ever computing the full gradient.

5 Numerical Experiments

In this section, we present numerical experiments testing B-SAGA, B-SVRG, SARAH, and SARGE for minimizing convex, strongly convex, and non-convex objectives. We include one set of experiments comparing different values of θ in B-SAGA and B-SVRG with a fixed step size and one set comparing SARAH and SARGE to B-SAGA and B-SVRG with the best values of θ.

5.1 Convex and strongly convex objectives

Let $(h_i, l_i) \in \mathbb{R}^p \times \{\pm 1\},\ i = 1, \cdots, n$ be the training set, where $h_i \in \mathbb{R}^p$ is the feature vector of each data sample, and l_i is the binary label. Let $\beta > 0$ be a tuning parameter. The ridge regression problem takes the form

$$\min_{x\in\mathbb{R}^p} \quad \frac{1}{n}\sum_{i=1}^n (h_i^\top x - l_i)^2 + \frac{\beta}{2}\|x\|_2^2.$$

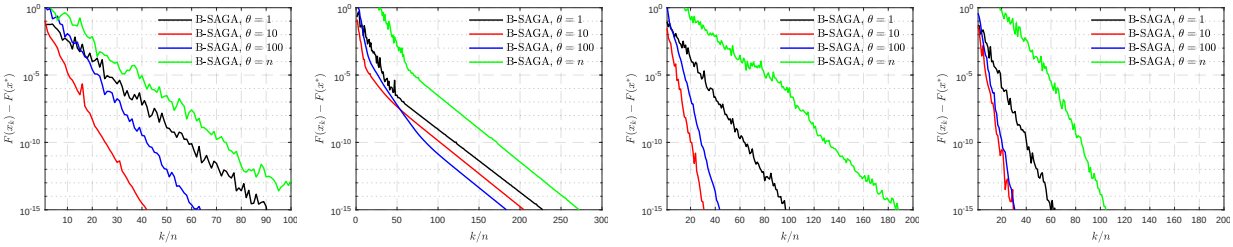
LASSO is similar, but with the regulariser $\|x\|_1$ replacing $\|x\|_2^2$. These problems are of the form (1), where we set $f_i = (h_i^\top x - l_i)^2$ and g equal to the regulariser. In ridge regression, g is strongly convex, and in LASSO, g is only convex.

We consider four binary classification data sets: `australian`, `mushrooms`, `phishing`, and `ijcnn1` from LIBSVM[2]. We rescale the value of the data to $[-1, 1]$, set $\beta = 1/n$, and set the step size to $\eta = \frac{1}{5L}$. To compare performance, we use the objective function value $F(x_k) - F(x^\star)$ is considered.

Comparison of B-SAGA We first compare the performance of B-SAGA under different choices of θ for solving ridge regression and LASSO problems. Four choices of θ are considered: $\theta \in \{1, 10, 100, n\}$, the results are provided below in Figures 1 and 2, from which we observe that B-SAGA consistently performs better with moderate amounts of bias (i.e. $\theta \in (1, n)$). For the considered datasets, overall $\theta = 10$ provides the best performance.

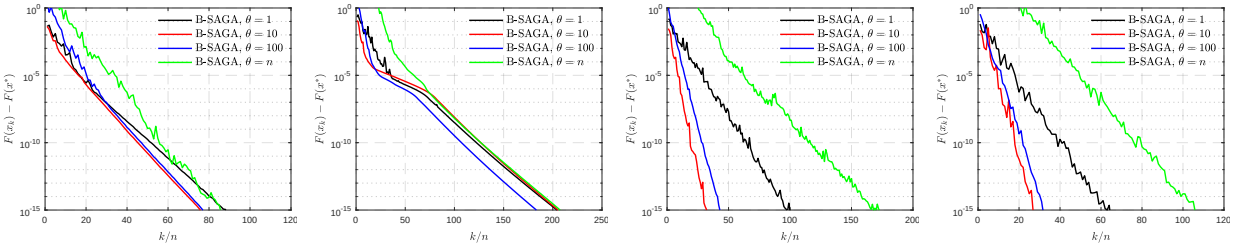
Comparison of B-SVRG We also consider four choices of θ for B-SVRG, which are $\theta \in \{0.5, 0.8, 1, 1.5\}$. The results are shown below in Figure 3 and 4. We observe that B-SVRG is more sensitive to the choice of θ; only small amounts of bias (i.e. $\theta \in [0.8, 1.5]$) can occasionally improve performance.

[2] `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`



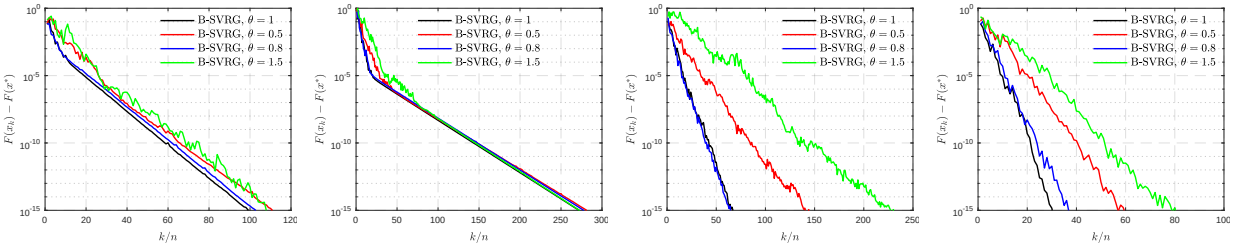
(a) `australian` (b) `mushrooms` (c) `phishing` (d) `ijcnn1`

Figure 1: Performance comparison fitting a ridge regression model for different choices of θ in B-SAGA. The step size for each case is set to $\eta = \frac{1}{5L}$.



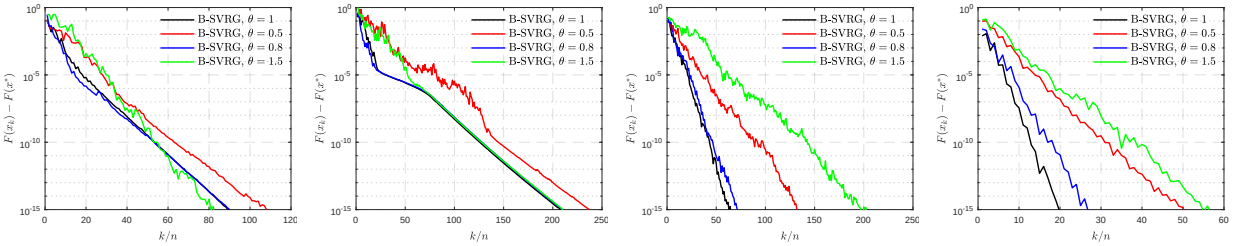
(a) `australian` (b) `mushrooms` (c) `phishing` (d) `ijcnn1`

Figure 2: Performance comparison fitting a LASSO model for different choices of θ in B-SAGA. The step size for each case is set to $\eta = \frac{1}{5L}$.



(a) `australian` (b) `mushrooms` (c) `phishing` (d) `ijcnn1`

Figure 3: Performance comparison fitting a ridge regression model for different choices of θ in B-SVRG. The step size for each case is set to $\eta = \frac{1}{5L}$.



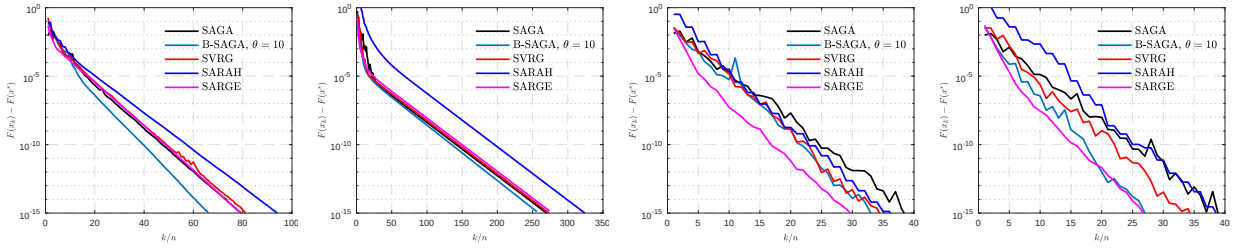
(a) `australian` (b) `mushrooms` (c) `phishing` (d) `ijcnn1`

Figure 4: Performance comparison fitting a LASSO model for different choices of θ in B-SVRG. The step size for each case is set to $\eta = \frac{1}{5L}$.

Comparison of different gradient estimators Finally, we provide comparison of SAGA, B-SAGA with $\theta = 10$, SVRG, SARAH and SARGE, the results are provided below in Figure 5 and 6 from which we observe that

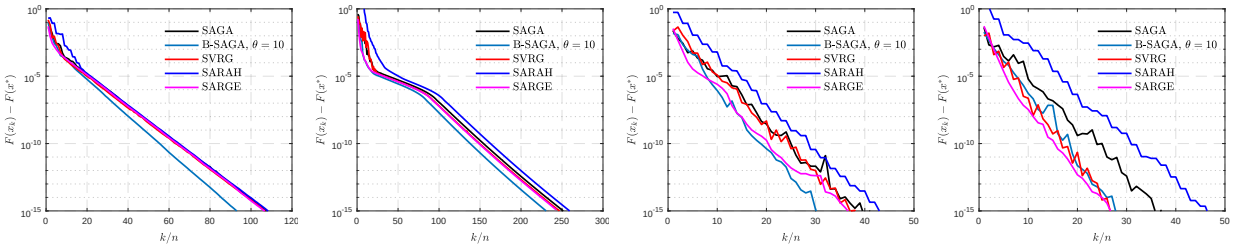
- SARAH performs similarly to SVRG, but is occasionally slower in early epochs.
- SARGE consistently outperforms all other methods except for B-SAGA with $\theta = 10$.

The above observations indicate that, depending on the data, biased schemes can benefit from their biased gradient estimates. The free parameter θ reduces the MSE of the B-SAGA and B-SVRG gradient estimators leading to better performance, and the bias in SARAH and SARGE has a similar effect.



(a) `australian` (b) `mushrooms` (c) `phishing` (d) `ijcnn1`

Figure 5: Performance comparison for solving ridge regression among different algorithms. Step sizes are tuned automatically to minimize the number of iterations required to reach a tolerance of 10^{-15}.



(a) `australian` (b) `mushrooms` (c) `phishing` (d) `ijcnn1`

Figure 6: Performance comparison for solving LASSO regression among different algorithms. Step sizes are tuned automatically to minimize the number of iterations required to reach a suboptimality of 10^{-15}.

5.2 Non-convex objectives

To test the effect of bias in the non-convex setting, we consider the non-negative principal component analysis (NN-PCA) problems, which can be formulated as [27]:

$$\min_{x \in \mathbb{R}^p} \quad -\tfrac{1}{n} \sum_{i=1}^n (h_i^\top x)^2 + \iota_C(x),$$

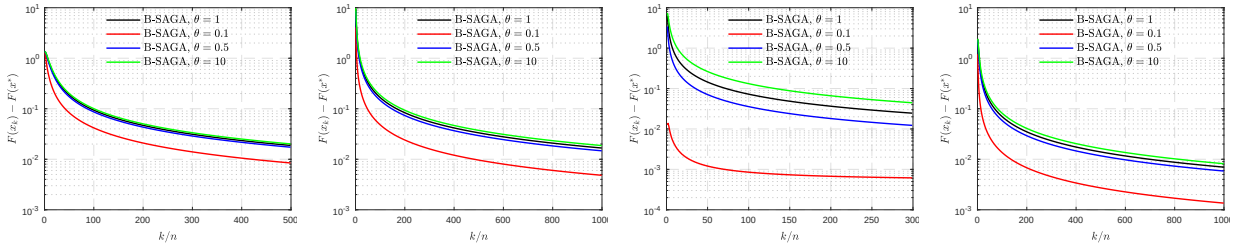
where $C \overset{\text{def}}{=} \{x \in \mathbb{R}^p : \|x\| \leq 1,\ x \geq 0\}$ is a convex set and

$$\iota_C(x) = \begin{cases} 0 : x \in C \\ +\infty : x \notin C \end{cases}$$

is the indicator function of C. Letting $g = \iota_C$, the operator $\text{prox}_{\eta g}$ is the projection onto C, which can be computed efficiently.

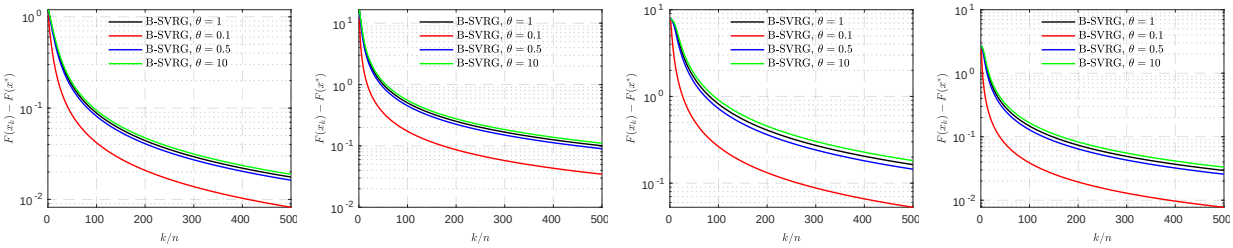
As the problem is non-convex, we cannot measure convergence with respect to the global optimum $x^\star$, so we use many iterations of proximal gradient descent with a small step size ($\eta = \frac{1}{10Ln}$) to find a reference

point $x^\star$. Every test is initialized using a random vector with normally distributed i.i.d. entries, and the same starting point is used for testing each value of θ. We found that small step sizes generally lead to stationary points with smaller objective values, so we set $\eta = \frac{1}{5n}$ for all our experiments. We report $F(x_k) - F(x^\star)$ averaged over every n iterations. These experiments show that the performance of B-SAGA and B-SVRG varies significantly with θ, with smaller values leading to better performance. SARAH and SARGE perform similarly to SAGA and SVRG in these experiments, see Figure 7 and 8.



(a) `australian` (b) `mushrooms` (c) `phishing` (d) `ijcnn1`

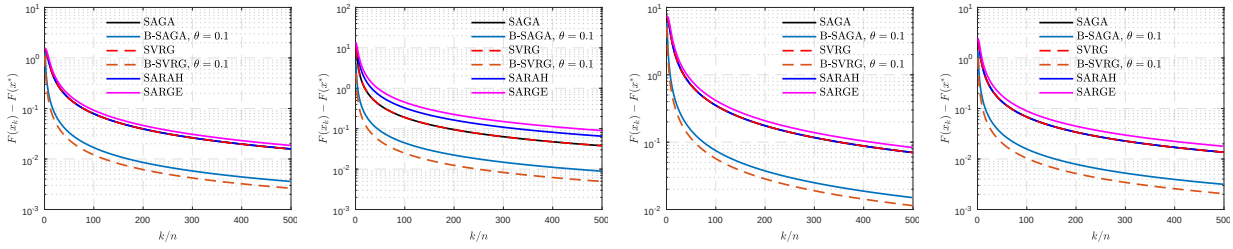
Figure 7: Performance comparison for solving NN-PCA with different choices of θ in B-SAGA. The step size for each case is set to $\eta = \frac{1}{5Ln}$. The point $x^\star$ is found by solving the problem using proximal gradient descent to high accuracy.



(a) `australian` (b) `mushrooms` (c) `phishing` (d) `ijcnn1`

Figure 8: Performance comparison for solving NN-PCA with different choices of θ in B-SVRG. The step size for each case is set to $\eta = \frac{1}{5Ln}$. The point $x^\star$ is found by solving the problem using proximal gradient descent.

For the comparison of all algorithms, B-SAGA and B-SVRG provides the best performance with B-SVRG being slightly faster.



(a) `australian` (b) `mushrooms` (c) `phishing` (d) `ijcnn1`

Figure 9: Performance comparison for solving NN-PCA among different algorithms. All step sizes are set to $\frac{1}{5Ln}$. Objective values are averaged over each epoch (n steps).

6 Conclusion

The complicated convergence proofs of biased stochastic gradient methods have restricted researchers to studying unbiased estimators almost exclusively. Our simple framework for proving convergence rates for biased algorithms overcomes this limitation. Our analysis allows for the study of biased algorithms with proximal support for minimizing convex, strongly convex, and non-convex objectives for the first time.

We also show that biased gradient estimators can offer improvements over unbiased estimators in theory and in practice. The B-SAGA and B-SVRG gradient estimators incorporate bias to reduce their mean squared errors and improve their performance in many settings. The bias in recursive gradient estimators, such as SARAH and SARGE, lead to much smaller bounds on their MSE's and faster convergence rates than B-SAGA and B-SVRG.

Acknowledgements

CBS acknowledges support from the Leverhulme Trust project on Breaking the Non-Convexity Barrier and on Unveiling the Invisible, the Philip Leverhulme Prize, the EPSRC grant No. EP/M00483X/1, the EPSRC Centre No. EP/N014588/1, the European Union Horizon 2020 research and innovation programmes under the Marie Skłodowska-Curie grant agreement No. 691070 CHiPS and the Marie Skłodowska-Curie grant agreement No 777826, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute.

References

[1] Allen-Zhu, Z. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. In *ICML* (2017).

[2] Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research* (2018), 1–51.

[3] Allen-Zhu, Z. Katyusha X: Practical momentum method for stochastic sum-of-nonconvex optimization. In *ICML* (2018).

[4] Allen-Zhu, Z. Natasha 2: Faster non-convex optimization than SGD. In 32^{nd} *Conference on Neural Information Processing Systems* (2018).

[5] Allen-Zhu, Z., and Hazan, E. Variance reduction for faster non-convex optimization. In *Proceedings of the* 33^{rd} *International Conference on Machine Learning* (2016), vol. 48.

[6] Allen-Zhu, Z., and Yuan, Y. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. In *ICML* (2018).

[7] Beck, A., and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences 2*, 1 (2009), 183–202.

[8] Bishop, C. M. *Pattern recognition and machine learning*. Springer, 2006.

[9] Bottou, L., Curtis, F. E., , and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review 60* (2018), 223–311.

[10] Bredies, K., and Lorenz, D. *Mathematical Image Processing*. Springer, 2018.

[11] Chambolle, A., Ehrhardt, M. J., Richtárik, P., and Schönlieb, C.-B. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM J. Optim. 28*, 4 (2018), 2783–2808.

[12] Defazio, A. A simple practical accelerated method for finite sums. In 30^{th} *Conference on Neural Information Processing Systems* (2016).

[13] Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems* (2014), pp. 1646–1654.

[14] Defazio, A., Caetano, T., and Domke, J. Finito: A faster, permutable incremental gradient method for big data problems. In *Proceedings of the 31st International Conference on Machine Learning* (2014).

[15] Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. In 32^{nd} *Conference on Neural Information Processing Systems* (2018).

[16] Garber, D., and Hazan, E. Faster and simple PCA via convex optimization. *arXiv:1509.05647v4* (2015).

[17] Hofmann, T., Lucchi, A., Lacoste-Julien, S., and McWilliams, B. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems* (2015), vol. 28, pp. 2296–2304.

[18] Johnson, R., and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems* (2013), pp. 315–323.

[19] Lan, G., Li, Z., and Zhou, Y. A unified variance-reduced accelerated gradient method for convex optimization. *arXiv:1905.12412* (2019).

[20] Liang, J., Fadili, J., and Peyré, G. Activity identification and local linear convergence of Forward–Backward-type methods. *SIAM Journal on Optimization 27*, 1 (2017), 408–437.

[21] Lions, P. L., and Mercier, B. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis 16*, 6 (1979), 964–979.

[22] Mairal, J. Incremental majorization-minimization optimization with application to large-scale machine learning. *Technical report* (2014).

[23] Nesterov, Y. *Introductory lectures on convex programming.* Springer, 2004.

[24] Nguyen, L. M., Liu, J., Scheinberg, K., and TakÁĈ, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning* (2017), vol. 70, pp. 2613–2621.

[25] Pham, N. H., Nguyen, L. M., Phan, D. T., and Tran-Dinh, Q. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *arXiv:1902.05679* (2019).

[26] Reddi, S. J., Hefny, A., Sra, S., Póczos, B., and Smola, A. Stochastic variance reduction for nonconvex optimization. In *Proc. 33rd International Conference on Machine Learning* (2016).

[27] Reddi, S. J., Sra, S., Póczos, B., and Smola, A. Fast stochastic methods for nonsmooth nonconvex optimization. In *Proc. 30th Annual Conference on Neural Information Processing Systems* (2016).

[28] Robbins, H., and Monro, S. A stochastic approximation method. *Annals of Mathematical Statistics 22*, 3 (1951), 400–407.

[29] Roux, N. L., Schmidt, M., and Bach, F. R. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in neural information processing systems* (2012), pp. 2663–2671.

[30] Schmidt, M., Roux, N. L., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming 162* (2017), 83–112.

[31] Shalev-Shwartz, S., and Zhang, T. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research 14* (2013), 567–599.

[32] Shang, F., Jiao, L., Zhou, K., Cheng, J., Ren, Y., and Jin, Y. ASVRG: Accelerated proximal SVRG. In *Asian Conference on Machine Learning* (2018), vol. 95, pp. 1–32.

[33] Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. SpiderBoost: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv:1810.10690* (2018).

[34] Xiao, L., and Zhang, T. A proximal stochastic gradient method with progressive variance reduction. *Technical report, Microsoft Research* (2014).

[35] Zhou, K., Shang, F., and Cheng, J. A simple stochastic variance reduced algorithm with fast convergence rates. In *ICML* (2018), pp. 5975–5984.

[36] Zhou, Y., Wang, Z., Ji, K., Liang, Y., and Tarokh, V. Momentum schemes with stochastic variance reduction for nonconvex composite optimization. *arXiv:1902.02715* (2019).

Appendix

The organization of the appendix is as follows: we prove Theorems 9 and 10 in Appendices A and B, respectively, and we prove Theorem 11 in Appendix C. We provide convergence rates for B-SAGA and B-SVRG as special cases of Theorem 9 in Appendix D, and we provide convergence rates for SARAH and SARGE as special cases of Theorem 10 in Appendices E and F, respectively.

A Proof of Theorem 9

To prove Theorem 9, we begin by showing that the BMSE property (Definition 2) implies that the MSE of the gradient estimator over T iterations is proportional to $\sum_{k=0}^{T-1}\mathbb{E}\|x_{k+1}-x_k\|^2$.

Lemma 20 (MSE bound) *Suppose that the stochastic gradient estimator $\widetilde{\nabla}$ satisfies the* BMSE$(M_1, M_2, \rho_M, \rho_F, m)$ *property, let $\rho = \min\{\rho_M, \rho_F\}$, and let σ_s be any sequence satisfying $\sigma_s(1-\rho)^{ms} \leq \sigma_{s-1}(1-\frac{\rho}{2})^{ms}$. For convenience, define $\Theta = \frac{M_1\rho_F+2M_2}{\rho_M\rho_F}$. The MSE of the gradient estimator is bounded as*

$$\sum_{s=0}^{S}\sigma_s\sum_{k=ms}^{m(s+1)-1}\mathbb{E}[\|\nabla f(x_k)-\widetilde{\nabla}_k\|^2] \leq 2\Theta L^2\sum_{s=0}^{S}\sigma_s\sum_{k=ms}^{m(s+1)-1}\mathbb{E}[\|x_{k+1}-x_k\|^2].$$

Proof. First, we derive a bound on the sequence $\mathcal{F}_{ms}$ arising in the BMSE property. Summing this sequence from $s=0$ to $s=S$,

$$\begin{aligned}
\sum_{s=0}^{S}\sigma_s\mathcal{F}_{ms} &\leq \sum_{s=0}^{S}\sum_{\ell=0}^{s}\tfrac{M_2\sigma_s(1-\rho_F)^{m(s-\ell)}}{n}\sum_{k=ms}^{m(s+1)-1}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2] \\
&\overset{①}{\leq} \sum_{s=0}^{S}\sum_{\ell=0}^{s}\tfrac{M_2\sigma_\ell(1-\frac{\rho_F}{2})^{m(s-\ell)}}{n}\sum_{k=ms}^{m(s+1)-1}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2] \\
&\leq \sum_{s=0}^{S}\left(\sum_{\ell=0}^{\infty}(1-\tfrac{\rho_F}{2})^{\ell}\right)\tfrac{M_2\sigma_s}{n}\sum_{k=ms}^{m(s+1)-1}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2] \\
&= \sum_{s=0}^{S}\tfrac{2M_2\sigma_s}{n\rho_F}\sum_{k=ms}^{m(s+1)-1}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2].
\end{aligned} \tag{8}$$

Inequality ① uses the fact that $\sigma_s(1-\rho_F)^{ms} \leq \sigma_{s-1}(1-\frac{\rho_F}{2})^{ms}$. With this bound on $\mathcal{F}_{ms}$, we proceed to

bound $\mathcal{M}_{ms}$ similarly.

$$
\begin{aligned}
\sum_{s=0}^{S}\sigma_s\mathcal{M}_{ms} &\overset{①}{\leq} \sum_{s=0}^{S}\sigma_s\Big(\mathcal{F}_{ms}+\tfrac{M_1}{n}\sum_{k=ms}^{m(s+1)-1}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2]\Big)+(1-\rho_M)^m\sum_{s=1}^{S}\sigma_s\mathcal{M}_{m(s-1)}\\
&\overset{②}{\leq} \sum_{s=0}^{S}\sigma_s\Big(\tfrac{M_1\rho_F+2M_2}{n\rho_F}\sum_{k=ms}^{m(s+1)-1}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2]\Big)+(1-\tfrac{\rho_M}{2})^m\sum_{s=1}^{S}\sigma_{s-1}\mathcal{M}_{m(s-1)}\\
&= \sum_{s=0}^{S}\sigma_s\Big(\tfrac{M_1\rho_F+2M_2}{n\rho_F}\sum_{k=ms}^{m(s+1)-1}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2]\Big)\\
&\quad+(1-\tfrac{\rho_M}{2})^m\sum_{s=1}^{S}\sigma_{s-1}\Big(\tfrac{M_1\rho_F+2M_2}{n\rho_F}\sum_{k=m(s-1)}^{ms-1}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2]\Big)+\cdots\\
&\leq \left(\sum_{\ell=0}^{\infty}(1-\tfrac{\rho_M}{2})^{m\ell}\right)\sum_{s=0}^{S}\sigma_s\Big(\tfrac{M_1\rho_F+2M_2}{n\rho_F}\sum_{k=ms}^{m(s+1)-1}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2]\Big)\\
&\overset{③}{\leq} \sum_{s=0}^{S}\tfrac{2\sigma_s\Theta}{n}\sum_{k=ms}^{m(s+1)-1}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2]\\
&\overset{④}{\leq} 2\Theta L^2\sum_{s=0}^{S}\sigma_s\sum_{k=ms}^{m(s+1)-1}\mathbb{E}[\|x_{k+1}-x_k\|^2].
\end{aligned}
$$

Inequality ① follows uses the fact that $\mathcal{M}_m \leq (1-\rho_M)^m\mathcal{M}_{m(s-1)}$. Inequality ② uses $\sigma_s(1-\rho_M)^{ms} \leq \sigma_{s-1}(1-\frac{\rho_M}{2})^{ms}$, ③ uses the same estimate we applied in (8), and ④ uses the Lipschitz continuity of ∇f_i. □

Proof of Lemma 7 By assumption, $1-\frac{1}{\theta}\geq 0$, so we can apply convexity to obtain

$$
\begin{aligned}
&\tfrac{\eta}{\theta}(f(x_k)-f(x^\star))+\tfrac{\eta}{n}(1-\tfrac{1}{\theta})\big(\textstyle\sum_{i=1}^{n}f_i(\varphi_k^i)-f_i(x^\star)\big)\\
&\leq \tfrac{\eta}{\theta}\langle\nabla f(x_k),x_k-x^\star\rangle+\tfrac{\eta}{n}(1-\tfrac{1}{\theta})\sum_{i=1}^{n}\langle\nabla f_i(\varphi_k^i),\varphi_k^i-x^\star\rangle\\
&= \tfrac{\eta}{\theta}\langle\nabla f(x_k),x_k-x^\star\rangle+\tfrac{\eta}{n}(1-\tfrac{1}{\theta})\sum_{i=1}^{n}\langle\nabla f_i(\varphi_k^i),x_k-x^\star\rangle+\tfrac{\eta}{n}(1-\tfrac{1}{\theta})\sum_{i=1}^{n}\langle\nabla f_i(\varphi_k^i),\varphi_k^i-x_k\rangle.
\end{aligned}
$$

Because $\widetilde{\nabla}_k$ is memory-biased,

$$
\tfrac{1}{\theta}\nabla f(x_k)+\tfrac{1}{n}(1-\tfrac{1}{\theta})\sum_{i=1}^{n}\nabla f_i(\varphi_k^i)=\mathbb{E}_k[\widetilde{\nabla}_k].
$$

Therefore,

$$
\begin{aligned}
&\tfrac{\eta}{\theta}\langle\nabla f(x_k),x_k-x^\star\rangle+\tfrac{\eta}{n}(1-\tfrac{1}{\theta})\sum_{i=1}^{n}\langle\nabla f_i(\varphi_k^i),x_k-x^\star\rangle\\
&= \mathbb{E}_k\big[\eta\langle\widetilde{\nabla}_k,x_k-x^\star\rangle\big]\\
&= \mathbb{E}_k[\eta\langle\widetilde{\nabla}_k,x_k-x_{k+1}\rangle+\eta\langle\widetilde{\nabla}_k,x_{k+1}-x^\star\rangle]\\
&\leq \mathbb{E}_k\big[\eta\langle\widetilde{\nabla}_k,x_k-x_{k+1}\rangle-\tfrac{1}{2}\|x_{k+1}-x_k\|^2+\tfrac{1}{2}\|x_k-x^\star\|^2-\tfrac{1+\mu\eta}{2}\|x_{k+1}-x^\star\|^2-\eta g(x_{k+1})+\eta g(x^\star)\big].
\end{aligned}
$$

The inequality is due to Lemma 3 with $z=x_{k+1}$, $x=x_k$, $d=\widetilde{\nabla}_k$, and $y=x^\star$. Combining these two

inequalities, we have shown

$$\begin{aligned}
&\tfrac{\eta}{\theta}(f(x_k)-f(x^\star))+\tfrac{\eta}{n}(1-\tfrac{1}{\theta})\sum_{i=1}^n\big(f_i(\varphi_k^i)-f_i(x^\star)\big)\\
&\le \mathbb{E}_k\big[\eta\langle\widetilde{\nabla}_k,x_k-x_{k+1}\rangle-\tfrac{1}{2}\|x_{k+1}-x_k\|^2-\eta g(x_{k+1})+\eta g(x^\star)\\
&\qquad+\tfrac{1}{2}\|x_k-x^\star\|^2-\tfrac{1+\mu\eta}{2}\|x_{k+1}-x^\star\|^2+\tfrac{\eta}{n}(1-\tfrac{1}{\theta})\sum_{i=1}^n\langle\nabla f_i(\varphi_k^i),\varphi_k^i-x_k\rangle\big].
\end{aligned}\tag{9}$$

We bound the first three terms on the right further.

$$\begin{aligned}
&\eta\langle\widetilde{\nabla}_k,x_k-x_{k+1}\rangle-\tfrac{1}{2}\|x_{k+1}-x_k\|^2-\eta g(x_{k+1})\\
&=\eta(\langle\nabla f(x_k),x_k-x_{k+1}\rangle-g(x_{k+1}))+\eta\langle\widetilde{\nabla}_k-\nabla f(x_k),x_k-x_{k+1}\rangle-\tfrac{1}{2}\|x_{k+1}-x_k\|^2\\
&\overset{①}{\le}\eta(f(x_k)-F(x_{k+1}))+\eta\langle\widetilde{\nabla}_k-\nabla f(x_k),x_k-x_{k+1}\rangle+(\tfrac{\eta L}{2}-\tfrac{1}{2})\|x_{k+1}-x_k\|^2\\
&\overset{②}{\le}\eta(f(x_k)-F(x_{k+1}))+\tfrac{\eta}{2L\lambda}\|\widetilde{\nabla}_k-\nabla f(x_k)\|^2+(\tfrac{\eta L(\lambda+1)}{2}-\tfrac{1}{2})\|x_{k+1}-x_k\|^2.
\end{aligned}$$

Inequality ① is due to the Lipschitz continuity of ∇f, and inequality ② is Young's. Combining this bound with (9) and rearranging terms, we have shown that

$$\begin{aligned}
0\le&-\eta\mathbb{E}_k[F(x_{k+1})-F(x^\star)]+\tfrac{\eta}{2L\lambda}\mathbb{E}_k[\|\widetilde{\nabla}_k-\nabla f(x_k)\|^2]\\
&-\tfrac{1+\mu\eta}{2}\mathbb{E}_k[\|x_{k+1}-x^\star\|^2]+\tfrac{1}{2}\|x_k-x^\star\|^2+(\tfrac{\eta L(\lambda+1)}{2}-\tfrac{1}{2})\mathbb{E}_k[\|x_{k+1}-x_k\|^2]\\
&+\eta(1-\tfrac{1}{\theta})\Big(f(x_k)-\tfrac{1}{n}\sum_{i=1}^n f_i(\varphi_k^i)+\tfrac{1}{n}\sum_{i=1}^n\langle\nabla f_i(\varphi_k^i),\varphi_k^i-x_k\rangle\Big).
\end{aligned}$$

We use Lemma 1 to bound the final term, yielding the desired inequality. ■

Proof of Theorem 9 (Convex Case) We begin with the inequality of Lemma 7 with $\mu=0$. Multiplying the inequality of Lemma 4 with $z=x_{k+1}$, $x=x_k$, and $d=\widetilde{\nabla}_k$ by a non-negative constant δ and adding it to the inequality of Lemma 7, we obtain

$$\begin{aligned}
&\eta\mathbb{E}_k[F(x_{k+1})-F(x^\star)+\delta(F(x_{k+1})-F(x_k))]\\
&\le\tfrac{\eta(1+\delta)}{2L\lambda}\mathbb{E}_k[\|\widetilde{\nabla}_k-\nabla f(x_k)\|^2]-\tfrac{1}{2}\mathbb{E}_k[\|x_{k+1}-x^\star\|^2]+\tfrac{1}{2}\|x_k-x^\star\|^2\\
&\qquad+(\tfrac{\eta L(1+\delta)(\lambda+1)}{2}-\tfrac{1+2\delta}{2})\mathbb{E}_k[\|x_{k+1}-x_k\|^2]+\tfrac{\eta L}{2n}(1-\tfrac{1}{\theta})\sum_{i=1}^n\|x_k-\varphi_k^i\|^2.
\end{aligned}$$

Applying the full expectation operator and summing from $k=0$ to $k=T-1$, we have

$$\begin{aligned}
&\eta\sum_{k=0}^{T-1}\mathbb{E}[F(x_{k+1})-F(x^\star)]+\eta\delta(\mathbb{E}[F(x_T)]-F(x_0))\\
&\le-\tfrac{1}{2}\mathbb{E}[\|x_T-x^\star\|^2]+\tfrac{1}{2}\|x_0-x^\star\|^2+\sum_{k=0}^{T-1}\mathbb{E}\big[\tfrac{\eta(1+\delta)}{2L\lambda}\|\widetilde{\nabla}_k-\nabla f(x_k)\|^2\\
&\qquad+(\tfrac{\eta L(1+\delta)(\lambda+1)}{2}-\tfrac{1+2\delta}{2})\|x_{k+1}-x_k\|^2+\tfrac{\eta L}{2n}(1-\tfrac{1}{\theta})\textstyle\sum_{i=1}^n\|x_k-\varphi_k^i\|^2\big].
\end{aligned}$$

We use Lemma 20 with $\sigma_s=1$ to bound the MSE, and we use the fact that the gradient estimator is memory-biased to bound the term $1/n\sum_{i=1}^n\|x_k-\varphi_k^i\|^2$. This leaves

$$\begin{aligned}
\eta\sum_{k=0}^{T-1}\mathbb{E}[F(x_{k+1})-F(x^\star)]\le&-\tfrac{1}{2}\mathbb{E}[\|x_T-x^\star\|^2]+\tfrac{1}{2}\|x_0-x^\star\|^2+\eta\delta(F(x_0)-\mathbb{E}[F(x_T)])\\
&+(\tfrac{\eta L(1+\delta)(\lambda+1)}{2}+\tfrac{\Theta\eta L(1+\delta)}{\lambda}+\tfrac{B_1\eta L}{2}(1-\tfrac{1}{\theta})-\tfrac{1+2\delta}{2})\sum_{k=0}^{T-1}\mathbb{E}[\|x_{k+1}-x_k\|^2].
\end{aligned}\tag{10}$$

Setting $\lambda = \sqrt{2\Theta}$ minimizes the coefficient of the term on the final line. With

$$\eta \leq \frac{1}{L(1+2\sqrt{2\Theta}+\frac{B_1(1-1/\theta)}{1+\delta})},$$

the final term in (10) is non-positive, so we can drop it from the inequality along with the term $-1/2\mathbb{E}\|x_T - x^\star\|^2$. Using the fact that $-F(x_T) \leq -F(x^\star)$, this leaves

$$\sum_{k=0}^{T-1} \mathbb{E}[F(x_{k+1}) - F(x^\star)] \leq \tfrac{1}{2\eta}\|x_0 - x^\star\|^2 + \eta\delta(F(x_0) - F(x^\star)).$$

We use the convexity of F to rewrite this inequality as a bound on the suboptimality of the average iterate

$$\mathbb{E}[F(\bar{x}_T) - F(x^\star)] \leq \frac{1}{T}\sum_{k=0}^{T-1} \mathbb{E}[F(x_{k+1}) - F(x^\star)] \leq \frac{1}{2\eta T}\|x_0 - x^\star\|^2 + \tfrac{\eta\delta}{T}(F(x_0) - F(x^\star)).$$

Setting $\delta = \max\{B_1(1-1/\theta)/\sqrt{2\Theta} - 1, 0\}$ approximately minimizes the right side, proving the assertion. ■

Proof of Theorem 9 (Strongly Convex Case) As in the proof of the convex case, we begin with the inequality of Lemma 7, multiply the inequality of Lemma 4 with $z = x_{k+1}$, $x = x_k$, and $d = \widetilde{\nabla}_k$ by a non-negative constant δ, and add the two inequalities.

$$\begin{aligned}
&\eta\mathbb{E}_k[F(x_{k+1}) - F(x^\star) + \delta(F(x_{k+1}) - F(x_k))] \\
&\leq -\tfrac{1+\mu\eta}{2}\mathbb{E}_k[\|x_{k+1} - x^\star\|^2] + \tfrac{1}{2}\|x_k - x^\star\|^2 + \mathbb{E}_k\Big[\tfrac{\eta(1+\delta)}{2L\lambda}\|\widetilde{\nabla}_k - \nabla f(x_k)\|^2 \\
&\quad + (\tfrac{\eta L(1+\delta)(\lambda+1)}{2} - \tfrac{1+\delta(2+\mu\eta)}{2})\|x_{k+1} - x_k\|^2 + \tfrac{\eta L}{2n}(1-\tfrac{1}{\theta})\textstyle\sum_{i=1}^n \|x_k - \varphi_k^i\|^2\Big].
\end{aligned}$$

Applying the full expectation operator, multiplying by $(1+\mu\eta)^k$, and summing over the epoch $k = ms$ to $k = m(s+1) - 1$ for some $s \in \mathbb{N}_0$, we have

$$\begin{aligned}
&\eta\sum_{k=ms}^{m(s+1)-1} (1+\mu\eta)^k\mathbb{E}[F(x_{k+1}) - F(x^\star) + \delta(F(x_{k+1}) - F(x_k))] \\
&\leq -\tfrac{(1+\mu\eta)^{m(s+1)}}{2}\mathbb{E}\|x_{m(s+1)} - x^\star\|^2 + \tfrac{(1+\mu\eta)^{ms}}{2}\mathbb{E}\|x_{ms} - x^\star\|^2 \\
&\quad + \sum_{k=ms}^{m(s+1)-1} (1+\mu\eta)^k\mathbb{E}\Big[\tfrac{\eta(1+\delta)}{2L\lambda}\|\widetilde{\nabla}_k - \nabla f(x_k)\|^2 + (\tfrac{\eta L(1+\delta)(\lambda+1)}{2} - \tfrac{1+\delta(2+\mu\eta)}{2})\|x_{k+1} - x_k\|^2 \\
&\quad + \tfrac{\eta L}{2n}(1-\tfrac{1}{\theta})\textstyle\sum_{i=1}^n \|x_k - \varphi_k^i\|^2\Big].
\end{aligned}$$

Using the fact that $\eta \leq \frac{1}{\mu m}$,

$$(1+\mu\eta)^k \leq (1+\mu\eta)^{m(s+1)} \leq (1+\mu\eta)^{ms}\lim_{m\to\infty}(1+\tfrac{1}{m})^m = e(1+\mu\eta)^{ms} \leq 3(1+\mu\eta)^{ms}, \tag{11}$$

where e is Euler's number. Therefore,

$$\begin{aligned}
&\eta\sum_{k=ms}^{m(s+1)-1} (1+\mu\eta)^k\mathbb{E}[F(x_{k+1}) - F(x^\star) + \delta(F(x_{k+1}) - F(x_k))] \\
&\leq -\tfrac{(1+\mu\eta)^{m(s+1)}}{2}\mathbb{E}[\|x_{m(s+1)} - x^\star\|^2] + \tfrac{(1+\mu\eta)^{ms}}{2}\|x_{ms} - x^\star\|^2 \\
&\quad + (1+\mu\eta)^{ms}\sum_{k=ms}^{m(s+1)-1} \mathbb{E}\Big[\tfrac{3\eta(1+\delta)}{2L\lambda}\|\widetilde{\nabla}_k - \nabla f(x_k)\|^2 + (\tfrac{3\eta L(1+\delta)(\lambda+1)}{2} - \tfrac{1+\delta(2+\mu\eta)}{2})\|x_{k+1} - x_k\|^2 \\
&\quad + \tfrac{3\eta L}{2n}(1-\tfrac{1}{\theta})\textstyle\sum_{i=1}^n \|x_k - \varphi_k^i\|^2\Big].
\end{aligned}$$

Summing the inequality from epoch $s = 0$ to $s = S - 1$,

$$\begin{aligned}
&\eta \sum_{k=0}^{mS-1} (1+\mu\eta)^k \mathbb{E}[F(x_{k+1}) - F(x^\star) + \delta(F(x_{k+1}) - F(x_k))] \\
&\leq \sum_{s=0}^{S-1} (1+\mu\eta)^{ms} \sum_{k=ms}^{m(s+1)-1} \mathbb{E}\Big[\tfrac{3\eta(1+\delta)}{2L\lambda} \|\widetilde{\nabla}_k - \nabla f(x_k)\|^2 + (\tfrac{3\eta L(1+\delta)(\lambda+1)}{2} - \tfrac{1+\delta(2+\mu\eta)}{2}) \|x_{k+1} - x_k\|^2 \\
&\quad + \tfrac{3\eta L(1+\delta)}{2n}(1 - \tfrac{1}{\theta}) \textstyle\sum_{i=1}^n \|x_k - \varphi_k^i\|^2 \Big] - \tfrac{(1+\mu\eta)^{mS}}{2} \mathbb{E}\|x_{ms} - x^\star\|^2 + \tfrac{1}{2}\|x_0 - x^\star\|^2.
\end{aligned}$$

We use Lemma 20 with $\sigma_s = (1+\mu\eta)^{ms}$ to bound the MSE. Recall $\rho = \min\{\rho_M, \rho_F\}$ and $\eta \leq \frac{\rho}{2\mu}$. This choice for σ_s satisfies the conditions of Lemma 20 because $(1+\mu\eta)^{ms}(1-\rho)^{ms} \leq (1+\mu\eta)^{m(s-1)}(1-\rho/2)^{ms}$. We use the fact that the gradient estimator is memory-biased to bound the term $1/n \sum_{i=1}^n \|x_k - \varphi_k^i\|^2$. This leaves

$$\begin{aligned}
&\eta \sum_{k=0}^{mS-1} (1+\mu\eta)^k \mathbb{E}[F(x_{k+1}) - F(x^\star) + \delta(F(x_{k+1}) - F(x_k))] \\
&\leq -\tfrac{(1+\mu\eta)^{mS}}{2} \mathbb{E}[\|x_{mS} - x^\star\|^2] + \tfrac{1}{2}\|x_0 - x^\star\|^2 + C \sum_{s=0}^{S-1} (1+\mu\eta)^{ms} \sum_{k=ms}^{m(s+1)-1} \mathbb{E}[\|x_{k+1} - x_k\|^2],
\end{aligned} \tag{12}$$

where $C = \frac{3\eta L(1+\delta)(\lambda+1)}{2} + \frac{3\Theta\eta L(1+\delta)}{\lambda} + \frac{3B_1\eta L}{2}(1 - \frac{1}{\theta}) - \frac{1+\delta(2+\mu\eta)}{2}$. We must choose η, λ, and δ so that $C \leq 0$. Setting $\lambda = \sqrt{2\Theta}$ minimizes C over λ. Using the approximation $\delta(2+\mu\eta) \geq \delta$, we see that C is non-positive if

$$\eta \leq \tfrac{1}{3L(1+2\sqrt{2\Theta} + \frac{B_1(1-1/\theta)}{1+\delta})}.$$

Setting $\delta = \max\{B_1(1-1/\theta)/\sqrt{2\Theta} - 1, 0\}$, we are guaranteed that

$$\tfrac{1}{3L(1+3\sqrt{2\Theta})} \leq \tfrac{1}{3L(1+2\sqrt{2\Theta} + \frac{B_1(1-1/\theta)}{1+\delta})},$$

so the step size in the theorem statement ensures $C \leq 0$, and the final term in (12) is non-positive. Dropping this non-positive term from the inequality, we have

$$\begin{aligned}
&\eta(1+\delta) \sum_{k=0}^{mS-1} (1+\mu\eta)^k \mathbb{E}[F(x_{k+1}) - F(x^\star)] + \delta\eta \sum_{k=0}^{mS-1} (1+\mu\eta)^k \mathbb{E}[F(x_k) - F(x^\star)] \\
&\leq -\tfrac{(1+\mu\eta)^{mS}}{2} \mathbb{E}[\|x_{mS} - x^\star\|^2] + \tfrac{1}{2}\|x_0 - x^\star\|^2.
\end{aligned} \tag{13}$$

We would like to show that $1 + \delta \geq (1+\mu\eta)\delta$ so that the terms on the first line telescope.

We use the fact that $\eta \leq \frac{\sqrt{2\Theta}}{B_1\mu(1-1/\theta)}$ to say

$$\tfrac{1}{\mu\eta} \geq \tfrac{B_1(1-1/\theta)}{\sqrt{2\Theta}} \geq \delta$$

Hence,

$$\tfrac{1+\delta}{\delta} \geq 1 + \mu\eta,$$

so inequality (13) simplifies to

$$(1+\mu\eta)^{mS} \mathbb{E}[\eta\delta(F(x_{mS}) - F(x^\star)) + \tfrac{1}{2}\|x_{mS} - x^\star\|^2] \leq \eta\delta(F(x_0) - F(x^\star)) + \tfrac{1}{2}\|x_0 - x^\star\|^2,$$

which implies the result. ■

B Proof of Theorem 10

The following two lemmas establish an analogue of Lemma 7 for recursively biased estimators.

Lemma 21 *Suppose $\widetilde{\nabla}$ is recursively biased with parameters ρ_B and ν. Suppose g is μ-strongly convex with $\mu \geq 0$, and let $\lambda > 0$ be a constant whose value we determine later. The following inequality holds:*

$$\begin{aligned}0 \leq &-\eta\mathbb{E}_k[F(x_{k+1}) - F(x^\star)] + \tfrac{\eta}{2L\lambda}\mathbb{E}_k[\|\widetilde{\nabla}_k - \nabla f(x_k)\|^2] \\ &- \tfrac{1+\mu\eta}{2}\mathbb{E}_k[\|x_{k+1} - x^\star\|^2] + \tfrac{1}{2}\|x_k - x^\star\|^2 \\ &+ (\tfrac{\eta L(\lambda+1)}{2} - \tfrac{1}{2})\mathbb{E}_k[\|x_{k+1} - x_k\|^2] + \eta(1-\rho_B)\langle \nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}, x_k - x^\star\rangle.\end{aligned}$$

Proof. Applying the convexity of f yields

$$\begin{aligned}&\eta(f(x_k) - f(x^\star)) \\ &\leq \eta\langle \nabla f(x_k), x_k - x^\star\rangle \\ &= \eta\langle \nabla f(x_k) - (1-\rho_B)(\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}), x_k - x^\star\rangle + \eta(1-\rho_B)\langle \nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}, x_k - x^\star\rangle.\end{aligned}$$

Because the estimator is recursively biased,

$$\mathbb{E}_k[\widetilde{\nabla}_k] = \nabla f(x_k) - (1-\rho_B)(\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}).$$

Therefore,

$$\begin{aligned}&\eta\langle \nabla f(x_k) - (1-\rho_B)(\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}), x_k - x^\star\rangle \\ &= \mathbb{E}_k[\eta\langle \widetilde{\nabla}_k, x_k - x^\star\rangle] \\ &= \mathbb{E}_k[\eta\langle \widetilde{\nabla}_k, x_k - x_{k+1}\rangle + \eta\langle \widetilde{\nabla}_k, x_{k+1} - x^\star\rangle] \\ &\leq \mathbb{E}_k[\eta\langle \widetilde{\nabla}_k, x_k - x_{k+1}\rangle - \tfrac{1}{2}\|x_{k+1} - x_k\|^2 + \tfrac{1}{2}\|x_k - x^\star\|^2 - \tfrac{1}{2}\|x_{k+1} - x^\star\|^2 + \eta g(x_{k+1}) - \eta g(x^\star)],\end{aligned}$$

The inequality is due to Lemma 3. The rest of the proof follows the proof of Lemma 7. □

Proof of Lemma 8 Because x_{k-1} is independent of j_{k-1}, we can use the BMSE property

$$\begin{aligned}&\mathbb{E}\langle \nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}, x_k - x^\star\rangle \\ &\overset{①}{=} \mathbb{E}[\langle \nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}, x_k - x_{k-1}\rangle + \langle \nabla f(x_{k-1}) - \mathbb{E}_{k-1}\widetilde{\nabla}_{k-1}, x_{k-1} - x^\star\rangle] \\ &\overset{②}{\leq} \mathbb{E}\big[\tfrac{\epsilon}{2}\|\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}\|^2 + \tfrac{1}{2\epsilon}\|x_k - x_{k-1}\|^2 + \langle \nabla f(x_{k-1}) - \mathbb{E}_{k-1}\widetilde{\nabla}_{k-1}, x_{k-1} - x^\star\rangle\big] \\ &\overset{③}{=} \mathbb{E}\big[\tfrac{\epsilon}{2}\|\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}\|^2 + \tfrac{1}{2\epsilon}\|x_k - x_{k-1}\|^2 + (1-\rho_B)\langle \nabla f(x_{k-2}) - \widetilde{\nabla}_{k-2}, x_{k-1} - x^\star\rangle\big].\end{aligned}$$

We can pass the conditional expectation $\mathbb{E}_{k-1}$ into the second inner-product in ① because x_{k-1} is independent of j_{k-1}. Inequality ② is Young's, and ③ uses the definition of a recursively biased gradient estimator.

This is a recursive inequality, and expanding the recursion gives

$$\begin{aligned}&\mathbb{E}\langle \nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}, x_k - x^\star\rangle \\ &\leq \sum_{\ell=\nu s+1}^{k-1}(1-\rho_B)^{k-\ell-1}\mathbb{E}\big[\tfrac{\epsilon}{2}\|\nabla f(x_\ell) - \widetilde{\nabla}_\ell\|^2 + \tfrac{1}{2\epsilon}\|x_{\ell+1} - x_\ell\|^2 + (1-\rho_B)\langle \nabla f(x_{\nu s}) - \widetilde{\nabla}_{\nu s}, x_{\nu s+1} - x^\star\rangle\big] \\ &\overset{①}{=} \sum_{\ell=\nu s+1}^{k-1}(1-\rho_B)^{k-\ell-1}\mathbb{E}\big[\tfrac{\epsilon}{2}\|\nabla f(x_\ell) - \widetilde{\nabla}_\ell\|^2 + \tfrac{1}{2\epsilon}\|x_{\ell+1} - x_\ell\|^2\big].\end{aligned}$$

Equality ① is due to the fact that $\widetilde{\nabla}_{\nu s} = \nabla f(x_{\nu s})$. Taking the absolute value and summing this from

$k = \nu s + 1$ to $k = \nu(s+1) - 1$,

$$\begin{aligned}
&\sum_{k=\nu s+1}^{\nu(s+1)-1} |\mathbb{E}\langle \nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}, x_k - x^\star\rangle| \\
&\leq \sum_{k=\nu s+1}^{\nu(s+1)-1} \sum_{\ell=\nu s+1}^{k-1} (1-\rho_B)^{k-\ell-1} \mathbb{E}\big[\tfrac{\epsilon}{2}\|\nabla f(x_\ell) - \widetilde{\nabla}_\ell\|^2 + \tfrac{1}{2\epsilon}\|x_{\ell+1} - x_\ell\|^2\big] \\
&\leq \min\Big\{\nu, \sum_{\ell=0}^{\infty}(1-\rho_B)^\ell\Big\} \sum_{k=\nu s+1}^{\nu(s+1)-1} \mathbb{E}\big[\tfrac{\epsilon}{2}\|\nabla f(x_k) - \widetilde{\nabla}_k\|^2 + \tfrac{1}{2\epsilon}\|x_{k+1} - x_k\|^2\big] \\
&\leq \min\Big\{\nu, \tfrac{1}{\rho_B}\Big\} \sum_{k=\nu s+1}^{\nu(s+1)-1} \mathbb{E}\big[\tfrac{\epsilon}{2}\|\nabla f(x_k) - \widetilde{\nabla}_k\|^2 + \tfrac{1}{2\epsilon}\|x_{k+1} - x_k\|^2\big].
\end{aligned}$$

Summing this inequality from $s = 0$ to $s = S$ completes the proof. ■

Proof of Theorem 10 (Convex Case) To begin, we sum the inequality of Lemma 21 and the inequality of Lemma 4 scaled by $\delta > 0$ with $z = x_{k+1}$, $x = x_k$, and $d = \widetilde{\nabla}_k$.

$$\begin{aligned}
&\eta\mathbb{E}_k[F(x_{k+1}) - F(x^\star) + \delta(F(x_{k+1}) - F(x_k))] \\
&\leq -\tfrac{1}{2}\mathbb{E}_k[\|x_{k+1} - x^\star\|^2] + \tfrac{1}{2}\|x_k - x^\star\|^2 + \mathbb{E}_k\Big[\tfrac{\eta(1+\delta)}{2L\lambda}\|\widetilde{\nabla}_k - \nabla f(x_k)\|^2 \\
&\qquad + (1+\delta)(\tfrac{\eta L(\lambda+1)}{2} - \tfrac{1}{2})\|x_{k+1} - x_k\|^2 + \eta(1-\rho_B)\langle \nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}, x_k - x^\star\rangle\Big].
\end{aligned} \tag{14}$$

Applying the full expectation operator, setting $\mu = 0$, and summing from $k = 0$ to $k = T-1$ where $T = mS$ for some $S \in \mathbb{N}$, we have

$$\begin{aligned}
&\eta\sum_{k=0}^{T-1}\mathbb{E}[F(x_{k+1}) - F(x^\star)] + \eta\delta\mathbb{E}[F(x_T) - F(x_0)] \\
&\leq -\tfrac{1}{2}\mathbb{E}[\|x_T - x^\star\|^2] + \tfrac{1}{2}\|x_0 - x^\star\|^2 + \sum_{k=0}^{T-1}\mathbb{E}\Big[\tfrac{\eta(1+\delta)}{2L\lambda}\|\widetilde{\nabla}_k - \nabla f(x_k)\|^2 \\
&\qquad + (1+\delta)(\tfrac{\eta L(\lambda+1)}{2} - \tfrac{1}{2})\|x_{k+1} - x_k\|^2 + \eta(1-\rho_B)\langle \nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}, x_k - x^\star\rangle\Big].
\end{aligned}$$

We use Lemma 8 to bound the inner-product bias term.

$$\begin{aligned}
&\eta\sum_{k=0}^{T-1}\mathbb{E}[F(x_{k+1}) - F(x^\star)] + \eta\delta\mathbb{E}[F(x_T) - F(x_0)] \\
&\leq -\tfrac{1}{2}\mathbb{E}[\|x_T - x^\star\|^2] + \tfrac{1}{2}\|x_0 - x^\star\|^2 + \sum_{k=0}^{T-1}\mathbb{E}\Big[(\tfrac{\eta(1+\delta)}{2L\lambda} + \tfrac{B_2\eta(1-\rho_B)\epsilon}{2})\|\widetilde{\nabla}_k - \nabla f(x_k)\|^2 \\
&\qquad + (1+\delta)(\tfrac{\eta L(\lambda+1)}{2} + \tfrac{B_2\eta(1-\rho_B)}{2\epsilon(1+\delta)} - \tfrac{1}{2})\|x_{k+1} - x_k\|^2\Big].
\end{aligned}$$

To bound the MSE, we use Lemma 20 with $\sigma_s = 1$. This leaves

$$\begin{aligned}
&\eta\sum_{k=0}^{T-1}\mathbb{E}[F(x_{k+1}) - F(x^\star)] + \eta\delta\mathbb{E}[F(x_T) - F(x_0)] \\
&\leq -\tfrac{1}{2}\mathbb{E}[\|x_T - x^\star\|^2] + \tfrac{1}{2}\|x_0 - x^\star\|^2 + w\sum_{k=0}^{T-1}\mathbb{E}[\|x_{k+1} - x_k\|^2],
\end{aligned} \tag{15}$$

where $w = \frac{\eta L(\lambda+1)(1+\delta)}{2} + \frac{B_2\eta(1-\rho_B)}{2\epsilon} + \frac{\Theta\eta L(1+\delta)}{\lambda} + B_2\eta L^2(1-\rho_B)\epsilon\Theta - \frac{1+\delta}{2}$. To minimize the coefficient of the final term, we set $\lambda = \sqrt{2\Theta}$ and $\epsilon = (2L^2\Theta)^{-1/2}$. This coefficient is then equal to

$$\sqrt{2\Theta}\eta L(1+\delta) + \tfrac{\eta L(1+\delta)}{2} + \sqrt{2}(1-\rho_B)\eta L B_2\sqrt{\Theta} - \tfrac{1+\delta}{2},$$

which is non-positive when $\eta \leq \frac{1}{2\sqrt{2\Theta}L(1+\frac{(1-\rho_B)B_2}{1+\delta})+L}$. This ensures that the final term in (15) is non-positive, so we can drop it from the inequality along with the term $-1/2\mathbb{E}\|x_T - x^\star\|^2$. This leaves

$$\sum_{k=0}^{T-1} \mathbb{E}[F(x_{k+1}) - F(x^\star)] \leq \tfrac{1}{2\eta}\|x_0 - x^\star\|^2 + \delta\eta\mathbb{E}[F(x_0) - F(x_T)].$$

By the convexity of F and the fact that $-F(x_T) \leq -F(x^\star)$

$$\mathbb{E}[F(\bar{x}_T) - F(x^\star)] \leq \tfrac{1}{T}\sum_{k=0}^{T-1} \mathbb{E}[F(x_{k+1}) - F(x^\star)] \leq \tfrac{1}{2\eta T}\|x_0 - x^\star\|^2 + \tfrac{\delta\eta}{T}(F(x_0) - F(x^\star)).$$

Choosing $\delta = \max\{(1-\rho_B)B_2 - 1, 0\}$ approximately minimizes the right side of this inequality, completing the proof. ■

Proof of Theorem 10 (Strongly Convex Case) We begin with inequality (14), but without setting $\mu = 0$.

$$\begin{aligned}
&\eta(1+\delta)\mathbb{E}_k[F(x_{k+1}) - F(x^\star)] + \tfrac{1+\mu\eta}{2}\mathbb{E}_k\|x_{k+1} - x^\star\|^2 \\
&\leq \eta\delta(F(x_k) - F(x^\star)) + \tfrac{1}{2}\|x_k - x^\star\|^2 + \mathbb{E}_k\Big[\tfrac{\eta(1+\delta)}{2L\lambda}\|\widetilde{\nabla}_k - \nabla f(x_k)\|^2 \\
&\quad + (1+\delta)(\tfrac{\eta L(\lambda+1)}{2} - \tfrac{1}{2})\|x_{k+1} - x_k\|^2 + \eta(1-\rho_B)\langle\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}, x_k - x^\star\rangle\Big].
\end{aligned}$$

Applying the full expectation operator, multiplying by $(1+\mu\eta)^k$, and summing over the epoch $k = ms$ to $k = m(s+1) - 1$ for some $s \in \mathbb{N}_0$, we have

$$\begin{aligned}
&\eta(1+\delta)\sum_{k=ms}^{m(s+1)-1}(1+\mu\eta)^k\mathbb{E}[F(x_{k+1}) - F(x^\star)] + \tfrac{(1+\mu\eta)^{m(s+1)}}{2}\mathbb{E}[\|x_{m(s+1)} - x^\star\|^2] \\
&\leq \eta\delta\sum_{k=ms}^{m(s+1)-1}(1+\mu\eta)^k\mathbb{E}[F(x_k) - F(x^\star)] + \tfrac{(1+\mu\eta)^{ms}}{2}\mathbb{E}[\|x_{ms} - x^\star\|^2] \\
&\quad + \sum_{k=ms}^{m(s+1)-1}(1+\mu\eta)^k\mathbb{E}\Big[\tfrac{\eta(1+\delta)}{2L\lambda}\|\widetilde{\nabla}_k - \nabla f(x_k)\|^2 + (1+\delta)(\tfrac{\eta L(\lambda+1)}{2} - \tfrac{1}{2})\|x_{k+1} - x_k\|^2 \\
&\quad + \eta(1-\rho_B)\langle\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}, x_k - x^\star\rangle\Big].
\end{aligned}$$

We would like to bound the inner-product bias term using Lemma 8, and we can do this after some manipulation. Because $\eta \leq \frac{1}{\mu m}$, we have $(1+\mu\eta)^k \leq 3(1+\mu\eta)^{ms})$. Using the same estimate as in equation (11), we can say

$$\begin{aligned}
&\sum_{k=ms}^{m(s+1)-1}(1+\mu\eta)^k\mathbb{E}[\langle\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}, x_k - x^\star\rangle] \\
&\leq 3(1+\mu\eta)^{ms}\sum_{k=ms}^{m(s+1)-1}|\mathbb{E}[\langle\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}, x_k - x^\star\rangle]|,
\end{aligned}$$

We can also choose δ so that $1 + \delta \geq (1+\mu\eta)\delta$. These simplifications lead to the inequality

$$\begin{aligned}
&(1+\mu\eta)^{m(s+1)}\mathbb{E}[\delta\eta(F(x_{m(s+1)}) - F(x^\star)) + \tfrac{1}{2}\|x_{m(s+1)} - x^\star\|^2] \\
&\leq \delta\eta(1+\mu\eta)^{ms}\mathbb{E}[F(x_{ms}) - F(x^\star)] + \tfrac{(1+\mu\eta)^{ms}}{2}\mathbb{E}\|x_{ms} - x^\star\|^2 \\
&\quad + (1+\mu\eta)^{ms}\Bigg(\sum_{k=ms}^{m(s+1)-1}\mathbb{E}\Big[\tfrac{3\eta(1+\delta)}{2L\lambda}\|\widetilde{\nabla}_k - \nabla f(x_k)\|^2 + (1+\delta)(\tfrac{3\eta L(\lambda+1)}{2} - \tfrac{1}{2})\|x_{k+1} - x_k\|^2\Big] \\
&\quad + 3\eta(1-\rho_B)|\mathbb{E}\langle\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}, x_k - x^\star\rangle|\Bigg).
\end{aligned}$$

Summing this inequality from $s = 0$ to $s = S - 1$,

$$
\begin{aligned}
&(1+\mu\eta)^{mS}\mathbb{E}[\delta\eta(F(x_{mS}) - F(x^\star)) + \tfrac{1}{2}\|x_{mS} - x^\star\|^2] \\
&\leq \delta\eta(F(x_0) - F(x^\star)) + \tfrac{1}{2}\|x_0 - x^\star\|^2 \\
&\quad + \sum_{s=0}^{S-1}(1+\mu\eta)^{ms}\Bigg(\sum_{k=ms}^{m(s+1)-1}\mathbb{E}\Big[\tfrac{3\eta(1+\delta)}{2L\lambda}\|\widetilde{\nabla}_k - \nabla f(x_k)\|^2 + (1+\delta)(\tfrac{3\eta L(\lambda+1)}{2} - \tfrac{1}{2})\|x_{k+1} - x_k\|^2\Big] \\
&\quad + 3\eta(1-\rho_B)|\mathbb{E}\langle\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}, x_k - x^\star\rangle|\Bigg).
\end{aligned}
$$

We use Lemma 20 with $\sigma_s = (1+\mu\eta)^{ms}$ to bound the MSE and Lemma 8 to bound the inner-product bias term.

$$
\begin{aligned}
&(1+\mu\eta)^{mS}\mathbb{E}[\delta\eta(F(x_{mS}) - F(x^\star)) + \tfrac{1}{2}\|x_{mS} - x^\star\|^2] \\
&\leq \delta\eta(F(x_0) - F(x^\star)) + \tfrac{1}{2}\|x_0 - x^\star\|^2 + w\sum_{s=0}^{S-1}\sum_{k=ms}^{m(s+1)-1}(1+\mu\eta)^{ms}\mathbb{E}\|x_{k+1} - x_k\|^2,
\end{aligned} \tag{16}
$$

where $w = \frac{3\eta L(\lambda+1)(1+\delta)}{2} + \frac{3B_2\eta(1-\rho_B)}{2\epsilon} + \frac{3\Theta\eta L(1+\delta)}{\lambda} + 3B_2\eta L^2(1-\rho_B)\epsilon\Theta - \frac{1+\delta}{2}$. To minimize the coefficient of the final term, we set $\lambda = \sqrt{2\Theta}$ and $\epsilon = (2L^2\Theta)^{-1/2}$. This coefficient is then equal to

$$
3\sqrt{2\Theta}\eta L(1+\delta) + \tfrac{3\eta L(1+\delta)}{2} + 3\sqrt{2}(1-\rho_B)\eta L B_2\sqrt{\Theta} - \tfrac{1+\delta}{2}.
$$

With

$$
\eta \leq \tfrac{1}{6\sqrt{2\Theta}L(1+\frac{(1-\rho_B)B_2}{1+\delta})+L}
$$

this term is non-positive. Setting $\delta = \max\{(1-\rho_B)B_2 - 1, 0\}$, we are assured that

$$
\eta \leq \tfrac{1}{3L(1+4\sqrt{2\Theta})} \leq \tfrac{1}{6\sqrt{2\Theta}L(1+\frac{(1-\rho_B)B_2}{1+\delta})+L},
$$

so the final term in (16) is non-positive, and we can drop it from the inequality. The resulting inequality is

$$
(1+\mu\eta)^T\mathbb{E}[\delta\eta(F(x_T) - F(x^\star)) + \tfrac{1}{2}\|x_T - x^\star\|^2] \leq \delta\eta(F(x_0) - F(x^\star)) + \tfrac{1}{2}\|x_0 - x^\star\|^2.
$$

All that remains is to show that our choice for δ satisfies $(1+\delta) \geq (1+\mu\eta)\delta$. Using the fact that

$$
\eta \leq \tfrac{1}{(1-\rho_B)B_2\mu},
$$

we can say

$$
\tfrac{1}{\mu\eta} \geq (1-\rho_B)B_2 \geq \delta.
$$

This ensures that $(1+\delta) \geq (1+\mu\eta)\delta$ and concludes the proof. ■

C Proof of Theorem 11

Theorem 11 follows immediately from inequality (6) and the MSE bound of Lemma 20.

Proof of Theorem 11

Summing inequality (6) from $k = 0$ to $k = T - 1$ and applying the full expectation operator, we obtain

$$
\begin{aligned}
0 \leq -\mathbb{E}[F(x_T)] + F(x_0) + (L - \tfrac{1}{4\eta})\sum_{k=0}^{T-1}\mathbb{E}[\|\hat{x}_{k+1} - x_k\|^2] \\
+ (\tfrac{L}{2} - \tfrac{1}{4\eta})\sum_{k=0}^{T-1}\mathbb{E}[\|x_{k+1} - x_k\|^2] + 2\eta\sum_{k=0}^{T-1}\mathbb{E}[\|\nabla f(x_k) - \widetilde{\nabla}_k\|^2].
\end{aligned}
$$

We bound the MSE using Lemma 20 with $\sigma_s = 1$.

$$0 \leq -\mathbb{E}[F(x_T)] + F(x_0) + (L - \tfrac{1}{4\eta}) \sum_{k=0}^{T-1} \mathbb{E}\|\hat{x}_{k+1} - x_k\|^2 + (\tfrac{L}{2} + 4\Theta\eta L^2 - \tfrac{1}{4\eta}) \sum_{k=0}^{T-1} \mathbb{E}\|x_{k+1} - x_k\|^2.$$

With $\eta \leq \frac{\sqrt{16\Theta+1}-1}{16L\Theta}$, the final term is non-positive, so we can drop it from the inequality. Using the fact that $-F(x_T) \leq -F(x^\star)$, our inequality simplifies to

$$-(L - \tfrac{1}{4\eta}) \sum_{k=0}^{T-1} \mathbb{E}[\|\hat{x}_{k+1} - x_k\|^2] \leq F(x_0) - F(x^\star).$$

Writing the left side in terms of the generalized gradient, we have the bound

$$\sum_{k=0}^{T-1} \mathbb{E}[\|\mathcal{G}_{\eta/2}(x_k)\|^2] \leq \frac{16(F(x_0) - F(x^\star))}{\eta(1 - 4\eta L)}.$$

With x_α chosen uniformly at random from the set $\{x_k\}_{k=0}^{T-1}$, this is equivalent to

$$\mathbb{E}[\|\mathcal{G}_{\eta/2}(x_\alpha)\|^2] \leq \frac{16(F(x_0) - F(x^\star))}{\eta(1 - 4\eta L)T}.$$

This completes the proof. ■

D Proofs of convergence rates for B-SAGA and B-SVRG

The following lemma establishes an MSE bound on the B-SAGA and B-SVRG gradient estimators. For the unbiased case $\theta = 1$, this result was essentially first proved in [13], but the authors ultimately use a looser variance bound.

Lemma 22 *The MSE's of the B-SAGA and B-SVRG gradient estimators satisfy*

$$\mathbb{E}_k[\|\widetilde{\nabla}_k - \nabla f(x_k)\|^2] \leq \frac{1}{n\theta^2} \sum_{i=1}^{n} \|\nabla f_i(x_k) - \nabla f_i(\varphi_k^i)\|^2 + (1 - \tfrac{2}{\theta})\|\nabla f(x_k) - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \nabla f_i(\varphi_k^i)\|^2. \tag{17}$$

Proof. Let $\widetilde{\nabla}_k \equiv \widetilde{\nabla}_k^{\text{B-SAGA}}$ or $\widetilde{\nabla}_k^{\text{B-SVRG}}$. The proof amounts to computing the expectation of the estimator and applying the Lipschitz continuity of ∇f_i.

$$\begin{aligned}
\mathbb{E}_k[\|\widetilde{\nabla}_k - \nabla f(x_k)\|^2] &= \mathbb{E}_k[\|\tfrac{1}{\theta}(\nabla f_{j_k}(x_k) - \nabla f_{j_k}(\varphi_k^{j_k})) - \nabla f(x_k) - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \nabla f_i(\varphi_k^i)\|^2] \\
&= \frac{1}{\theta^2}\mathbb{E}_k[\|\nabla f_{j_k}(x_k) - \nabla f_{j_k}(\varphi_k^{j_k})\|^2] + \|\nabla f(x_k) - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \nabla f_i(\varphi_k^i)\|^2 \\
&\quad - \frac{2}{\theta}\mathbb{E}_k[\langle \nabla f_{j_k}(x_k) - \nabla f_{j_k}(\varphi_k^{j_k}), \nabla f(x_k) - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \nabla f_i(\varphi_k^i)\rangle] \\
&= \frac{1}{n\theta^2} \sum_{i=1}^{n} \|\nabla f_i(x_k) - \nabla f_i(\varphi_k^i)\|^2 + (1 - \tfrac{2}{\theta})\|\nabla f(x_k) - \frac{1}{n}\textstyle\sum_{i=1}^{n} \nabla f_i(\varphi_k^i)\|^2,
\end{aligned}$$

which is the desired result. □

The following two lemmas establish the constants in the BMSE property for the B-SAGA and B-SVRG estimators.

Proof of Lemma 12 We begin with the inequality of Lemma 22 and consider two cases.

Case 1. Suppose $\theta \in [1,2]$. In this case the second term in (17) is non-positive, so we drop it from the inequality. For the remaining term, we use the following bound.

$$\begin{aligned}
&\tfrac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_k)-\nabla f_i(\varphi_k^i)\|^2] \\
&\overset{①}{\leq} \tfrac{1+2n}{n}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_k)-\nabla f_i(x_{k-1})\|^2]+\tfrac{1}{n}(1+\tfrac{1}{2n})\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_{k-1})-\nabla f_i(\varphi_k^i)\|^2] \\
&\overset{②}{=} \tfrac{1+2n}{n}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_k)-\nabla f_i(x_{k-1})\|^2]+\tfrac{1}{n}(1+\tfrac{1}{2n})(1-\tfrac{1}{n})\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_{k-1})-\nabla f_i(\varphi_{k-1}^i)\|^2] \\
&\overset{③}{\leq} \tfrac{1+2n}{n}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_k)-\nabla f_i(x_{k-1})\|^2]+\tfrac{1}{n}(1-\tfrac{1}{2n})\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_{k-1})-\nabla f_i(\varphi_{k-1}^i)\|^2].
\end{aligned} \tag{18}$$

Inequality ① is the standard inequality $\|a-c\|^2 \leq (1+\delta)\|a-b\|^2+(1+\delta^{-1})\|b-c\|^2$ (where we let $\delta=\frac{1}{2n}$). Inequality ② follows from the definition of φ_k^i and computing the expectation over j_{k-1}, and ③ uses the fact that $(1+\frac{1}{2n})(1-\frac{1}{n}) \leq (1-\frac{1}{2n})$. Altogether, this gives

$$\begin{aligned}
&\mathbb{E}[\|\widetilde{\nabla}_k^{\mathrm{SAGA}}-\nabla f(x_k)\|^2] \\
&\leq \tfrac{1}{n\theta^2}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_k)-\nabla f_i(\varphi_k^i)\|] \\
&\leq \tfrac{2n+1}{n\theta^2}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_k)-\nabla f_i(x_{k-1})\|^2]+\tfrac{1}{n\theta^2}(1-\tfrac{1}{2n})\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_{k-1})-\nabla f_i(\varphi_{k-1}^i)\|^2].
\end{aligned}$$

With $\mathcal{M}_k = \frac{1}{n\theta^2}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_k)-\nabla f_i(\varphi_k^i)\|^2]$, it is clear that the SAGA estimator satisfies the BMSE property with $M_1=\frac{2n+1}{\theta^2}, \rho_M=\frac{1}{2n}, M_2=0, \rho_F=1$, and $m=1$.

Case 2. Suppose $\theta > 2$, so that the second term in (17) is non-negative. Jensen's inequality gives

$$\mathbb{E}_k[\|\widetilde{\nabla}_k-\nabla f(x_k)\|^2] \leq \tfrac{1}{n}(1-\tfrac{1}{\theta})^2\sum_{i=1}^{n}\|\nabla f_i(x_k)-\nabla f_i(\varphi_k^i)\|^2.$$

Following the argument of Case 1, it is easy to see that the B-SAGA gradient estimator satisfies the BMSE property with $\mathcal{M}_k = \frac{1}{n}(1-\frac{1}{\theta})^2\sum_{i=1}^{n}\|\nabla f_i(x_k)-\nabla f_i(\varphi_k^i)\|^2$, $M_1=(2n+1)(1-\frac{1}{\theta})^2$, $\rho_M=\frac{1}{2n}$, $M_2=0$, $\rho_F=1$, and $m=1$.

To prove that the B-SAGA estimator is memory-biased, we must only compute its expectation.

$$\begin{aligned}
\nabla f(x_k)-\mathbb{E}_k[\widetilde{\nabla}_k^{\text{B-SAGA}}] &= \nabla f(x_k)-\tfrac{1}{\theta}\mathbb{E}_k[\nabla f_{j_k}(x_k)-f_i(\varphi_k^{j_k})]-\tfrac{1}{n}\sum_{i=1}^{n}\nabla f_i(\varphi_k^i) \\
&= (1-\tfrac{1}{\theta})\Big(\nabla f(x_k)-\tfrac{1}{n}\sum_{i=1}^{n}\nabla f_i(\varphi_k^i)\Big).
\end{aligned}$$

To compute a value for B_1, we follow (18) to obtain

$$\begin{aligned}
\tfrac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|x_k-\varphi_k^i\|^2] &\leq (2n+1)\|x_k-x_{k-1}\|^2+\tfrac{1}{n}(1-\tfrac{1}{2n})\sum_{i=1}^{n}\mathbb{E}[\|x_{k-1}-\varphi_{k-1}^i\|^2] \\
&\leq (2n+1)\sum_{\ell=1}^{k}(1-\tfrac{1}{2n})^{k-\ell}\|x_\ell-x_{\ell-1}\|^2.
\end{aligned}$$

Summing this inequality from $k=0$ to $k=T-1$, we obtain

$$\begin{aligned}\frac{1}{n}\sum_{k=0}^{T-1}\sum_{i=1}^{n}\mathbb{E}[\|x_k-\varphi_k^i\|^2] &\le (2n+1)\sum_{k=0}^{T-1}\sum_{\ell=1}^{k}(1-\tfrac{1}{2n})^{k-\ell}\|x_\ell-x_{\ell-1}\|^2\\ &\le (2n+1)\big(\textstyle\sum_{\ell=0}^{\infty}(1-\frac{1}{2n})^{\ell}\big)\displaystyle\sum_{k=0}^{T-1}\|x_{k+1}-x_k\|^2\\ &= 2n(2n+1)\sum_{k=0}^{T-1}\|x_{k+1}-x_k\|^2,\end{aligned}$$

which completes the proof. ■

Proof of Lemma 13 Suppose $k\in\{ms, ms+1,\cdots m(s+1)-1\}$ for some $s\in\mathbb{N}_0$. As in the proof of Lemma 12, we begin with the inequality of Lemma 22 and consider two cases.

Case 1. Suppose $\theta\in[1,2]$, so that we may drop the second term in (17). We can bound the remaining term as follows.

$$\begin{aligned}\tfrac{1}{n\theta^2}\sum_{i=1}^{n}\|\nabla f_i(x_k)-\nabla f_i(\varphi_s)\|^2 &\overset{①}{\le} \tfrac{1+m}{n\theta^2}\sum_{i=1}^{n}\|\nabla f_i(x_k)-\nabla f_i(x_{k-1})\|^2+\tfrac{1+1/m}{n\theta^2}\sum_{i=1}^{n}\|\nabla f_i(x_{k-1})-\nabla f_i(\varphi_s)\|^2\\ &\overset{②}{\le} \tfrac{1+m}{n\theta^2}\sum_{\ell=ms}^{k}(1+\tfrac{1}{m})^{k-\ell}\sum_{i=1}^{n}\|\nabla f_i(x_{\ell+1})-\nabla f_i(x_\ell)\|^2.\end{aligned}$$

Inequality ① uses the inequality $\|u-w\|^2\le(1+1/m)\|u-v\|^2+(1+m)\|v-w\|^2$, and ② follows from the fact that $x_{ms}=\varphi_s$. Summing this inequality from $k=ms$ to $k=m(s+1)-1$ gives us

$$\begin{aligned}\tfrac{1}{n\theta^2}\sum_{k=ms}^{m(s+1)-1}\sum_{i=1}^{n}\|\nabla f_i(x_k)-\nabla f_i(\varphi_s)\|^2 &\le \tfrac{m+1}{n\theta^2}(1+\tfrac{1}{m})^m\sum_{k=ms}^{m(s+1)-1}\sum_{\ell=ms}^{k}\sum_{i=1}^{n}\|\nabla f_i(x_{\ell+1})-\nabla f_i(x_\ell)\|^2\\ &\le \tfrac{m(m+1)}{n\theta^2}(1+\tfrac{1}{m})^m\sum_{k=ms}^{m(s+1)-1}\sum_{i=1}^{n}\|\nabla f_i(x_{\ell+1})-\nabla f_i(x_\ell)\|^2\\ &\le \tfrac{3m(m+1)}{n\theta^2}\sum_{k=ms}^{m(s+1)-1}\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2.\end{aligned}$$

The final inequality uses the fact that $(1+\frac{1}{m})^m<\lim_{m\to\infty}(1+\frac{1}{m})^m=e<3$. From this inequality, it is clear that the B-SVRG gradient estimator satisfies the BMSE property with $M_1=\frac{3m(m+1)}{\theta^2}$, $\rho_M=1$, $M_2=0$, and $\rho_F=1$.

Case 2. If $\theta>2$, then applying Jensen's inequality to (17) produces

$$\mathbb{E}_k[\|\widetilde{\nabla}_k^{\text{B-SVRG}}-\nabla f(x_k)\|^2]\le\tfrac{1}{n}(1-\tfrac{1}{\theta})^2\sum_{i=1}^{n}\|\nabla f_i(x_k)-\nabla f_i(\varphi_s)\|^2.$$

A similar argument to the one in Case 1 shows that $M_1=3m(m+1)(1-\frac{1}{\theta})^2$, $\rho_M=1$, $M_2=0$, and $\rho_F=1$.

All that is left is to prove the stated value for B_1. Following the proof in Case 1,

$$\sum_{k=ms}^{m(s+1)-1}\|x_k-\varphi_s\|^2\le\sum_{k=ms}^{m(s+1)-1}\sum_{\ell=ms}^{k}(1+m)(1+\frac{1}{m})^m\|x_{\ell+1}-x_\ell\|^2\le 3m(m+1)\sum_{k=ms}^{m(s+1)-1}\|x_{k+1}-x_k\|^2.$$

Summing over the epochs $s=0$ to $s=S$ shows $B_1=3m(m+1)$. ■

Combining Lemmas 12 and 13 with Theorems 9 and 11 proves convergence rates for B-SAGA and B-SVRG.

E Proof of convergence rates for SARAH

Lemma 16 establishes the BMSE constants for the SARAH estimator. The convergence rates of Corollary 18 then follow immediately from Theorem 10.

Proof of Lemma 16 Let $k \in \{ms+1, ms+2, \cdots, m(s+1)-1\}$. The claim follows immediately from the well-known bound on the MSE of the SARAH gradient estimator

$$\|\widetilde{\nabla}_k^{\text{SARAH}} - \nabla f(x_k)\|^2 \leq \frac{1}{n} \sum_{\ell=ms}^{k} \sum_{i=1}^{n} \|\nabla f_i(x_{\ell+1}) - \nabla f_i(x_\ell)\|^2.$$

We refer to [15], for example, for a proof of this inequality. Summing over an epoch and applying the estimate

$$\frac{1}{n} \sum_{k=ms}^{m(s+1)-1} \sum_{\ell=ms}^{k} \sum_{i=1}^{n} \|\nabla f_i(x_{\ell+1}) - \nabla f_i(x_\ell)\|^2 \leq \frac{m}{n} \sum_{k=ms}^{m(s+1)-1} \sum_{i=1}^{n} \|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\|^2$$

complete the proof. ■

F Proof of convergence rates for SARGE

For our analysis, we write the SARGE gradient estimator in terms of the SAGA estimator. Define the estimator

$$\widetilde{\nabla}_k^{\xi\text{-SAGA}} \stackrel{\text{def}}{=} \nabla f_{j_k}(x_{k-1}) - \nabla f_{j_k}(\xi_k^{j_k}) + \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\xi_k^i),$$

where the variables $\{\xi_k^i\}_{i=1}^n$ follow the update rules $\xi_{k+1}^{j_k} = x_{k-1}$ and $\xi_{k+1}^i = \xi_k^i$ for all $i \neq j_k$. The SARGE estimator is equal to

$$\widetilde{\nabla}_k^{\text{SARGE}} = \widetilde{\nabla}_k^{\text{SAGA}} - (1 - \tfrac{1}{n})(\widetilde{\nabla}_k^{\xi\text{-SAGA}} - \widetilde{\nabla}_k^{\text{SARGE}}).$$

Before we prove Lemma 17, we require a bound on the MSE of the ξ-SAGA gradient estimator that follows immediately from Lemma 22.

Lemma 23 *The MSE of the ξ-SAGA gradient estimator satisfies the following bound:*

$$\mathbb{E}[\|\widetilde{\nabla}_k^{\xi\text{-SAGA}} - \nabla f(x_{k-1})\|^2] \leq 3 \sum_{\ell=1}^{k-1} (1 - \tfrac{1}{2n})^{k-\ell-1} \sum_{i=1}^{n} \mathbb{E}[\|\nabla f_i(x_\ell) - \nabla f_i(x_{\ell-1})\|^2].$$

Proof. Following the proof of Lemma 22,

$$\begin{aligned}
\mathbb{E}_k[\|\widetilde{\nabla}_k^{\xi\text{-SAGA}} - \nabla f(x_{k-1})\|^2] &= \mathbb{E}_k[\|\nabla f_{j_k}(x_{k-1}) - \nabla f_{j_k}(\xi_k^{j_k}) - \nabla f(x_{k-1}) + \tfrac{1}{n} \textstyle\sum_{i=1}^n \nabla f_i(\xi_k^i)\|^2] \\
&\overset{①}{=} \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x_{k-1}) - \nabla f_i(\xi_k^i)\|^2 - \|\nabla f(x_{k-1}) - \tfrac{1}{n} \textstyle\sum_{i=1}^n \nabla f_i(\xi_k^i)\|^2 \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x_{k-1}) - \nabla f_i(\xi_k^i)\|^2.
\end{aligned}$$

Equality ① is the standard variance decomposition. To continue, we follow the proof of Lemma 22.

$$
\begin{aligned}
&\mathbb{E}[\|\widetilde{\nabla}_k^{\xi\text{-SAGA}} - \nabla f(x_{k-1})\|^2] \\
&\leq \tfrac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\nabla f_i(x_{k-1}) - \nabla f_i(\xi_k^i)\|^2] \\
&\leq \tfrac{1+2n}{n}\sum_{i=1}^n \mathbb{E}[\|\nabla f_i(x_{k-1}) - \nabla f_i(x_{k-2})\|^2] + \tfrac{1}{n}(1+\tfrac{1}{2n})\sum_{i=1}^n \mathbb{E}[\|\nabla f_i(x_{k-2}) - \nabla f_i(\xi_k^i)\|^2] \\
&\overset{②}{=} \tfrac{(1+2n)}{n}\sum_{i=1}^n \mathbb{E}[\|\nabla f_i(x_{k-1}) - \nabla f_i(x_{k-2})\|^2] + \tfrac{1}{n}(1+\tfrac{1}{2n})(1-\tfrac{1}{n})\sum_{i=1}^n \mathbb{E}[\|\nabla f_i(x_{k-2}) - \nabla f_i(\xi_{k-1}^i)\|^2] \\
&\overset{③}{\leq} 3\sum_{i=1}^n \mathbb{E}[\|\nabla f_i(x_{k-1}) - \nabla f_i(x_{k-2})\|^2] + \tfrac{1}{n}(1-\tfrac{1}{2n})\sum_{i=1}^n \mathbb{E}[\|\nabla f_i(x_{k-2}) - \nabla f_i(\xi_{k-1}^i)\|^2] \\
&\leq 3\sum_{\ell=1}^{k-1}(1-\tfrac{1}{2n})^{k-\ell-1}\sum_{i=1}^n \mathbb{E}[\|\nabla f_i(x_\ell) - \nabla f_i(x_{\ell-1})\|^2].
\end{aligned}
$$

Equality ② follows from computing expectations, and ③ uses the estimate $(1-\frac{1}{n})(1+\frac{1}{2n}) \leq (1-\frac{1}{2n})$. □

Due to the recursive nature of the SARGE gradient estimator, its MSE depends on the difference between the current estimate and the estimate from the previous iteration. The next lemma provides a bound on $\|\widetilde{\nabla}_k^{\text{SARGE}} - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2$.

Lemma 24 *The SARGE gradient estimator satisfies the following bound:*

$$
\begin{aligned}
\mathbb{E}[\|\widetilde{\nabla}_k^{\text{SARGE}} - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2] \leq \tfrac{12}{n}\sum_{i=1}^n \mathbb{E}[\|\nabla f_i(x_k) - \nabla f_i(x_{k-1})\|^2] + \tfrac{3}{2n^2}\mathbb{E}\|\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2 \\
+ \tfrac{39}{n^2}\sum_{\ell=1}^k (1-\tfrac{1}{2n})^{k-\ell}\sum_{i=1}^n \mathbb{E}\|\nabla f_i(x_\ell) - \nabla f_i(x_{\ell-1})\|^2.
\end{aligned}
$$

Proof. To begin, we use the standard inequality $\|a-c\|^2 \leq (1+\delta)\|a-b\|^2 + (1+\delta^{-1})\|b-c\|^2$ for any $\delta > 0$ twice. For simplicity, we set $\delta = \sqrt{3/2} - 1$ and use the fact that $1 + \frac{1}{\sqrt{3/2}-1} \leq 6$ for both applications of this inequality.

$$
\begin{aligned}
&\mathbb{E}[\|\widetilde{\nabla}_k^{\text{SARGE}} - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2] \\
&= \mathbb{E}[\|\widetilde{\nabla}_k^{\text{SAGA}} - (1-\tfrac{1}{n})(\widetilde{\nabla}_k^{\xi\text{-SAGA}} - \widetilde{\nabla}_{k-1}^{\text{SARGE}}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2] \\
&\leq 6\mathbb{E}[\|\widetilde{\nabla}_k^{\text{SAGA}} - \widetilde{\nabla}_k^{\xi\text{-SAGA}}\|^2] + \tfrac{\sqrt{3}}{\sqrt{2}n^2}\mathbb{E}[\|\widetilde{\nabla}_k^{\xi\text{-SAGA}} - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2] \\
&\leq 6\mathbb{E}[\|\widetilde{\nabla}_k^{\text{SAGA}} - \widetilde{\nabla}_k^{\xi\text{-SAGA}}\|^2] + \tfrac{6\sqrt{3}}{\sqrt{2}n^2}\mathbb{E}[\|\widetilde{\nabla}_k^{\xi\text{-SAGA}} - \nabla f(x_{k-1})\|^2] + \tfrac{3}{2n^2}\mathbb{E}[\|\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2].
\end{aligned} \tag{19}
$$

We use $\frac{6\sqrt{3}}{\sqrt{2}n^2} \leq \frac{9}{n^2}$ to simplify the coefficient of the second term. We now bound the first two of these three terms separately. Consider the first term.

$$
\begin{aligned}
&6\mathbb{E}[\|\widetilde{\nabla}_k^{\text{SAGA}} - \widetilde{\nabla}_k^{\xi\text{-SAGA}}\|^2] \\
&= 6\mathbb{E}[\|\nabla f_{j_k}(x_k) - \nabla f_{j_k}(\varphi_k^{j_k}) + \tfrac{1}{n}\textstyle\sum_{i=1}^n \nabla f_i(\varphi_k^i) - \nabla f_{j_k}(x_{k-1}) - \nabla f_{j_k}(\xi_k^{j_k}) - \tfrac{1}{n}\sum_{i=1}^n \nabla f_i(\xi_k^i)\|^2] \\
&\leq 12\mathbb{E}[\|\nabla f_{j_k}(x_k) - \nabla f_{j_k}(x_{k-1})\|^2] \\
&\quad + 12\mathbb{E}[\|\nabla f_{j_k}(\varphi_k^{j_k}) - \nabla f_{j_k}(\xi_k^{j_k}) - \tfrac{1}{n}\textstyle\sum_{i=1}^n \nabla f_i(\varphi_k^i) + \tfrac{1}{n}\sum_{i=1}^n \nabla f_i(\xi_k^i)\|^2] \\
&\overset{①}{=} 12\mathbb{E}[\|\nabla f_{j_k}(x_k) - \nabla f_{j_k}(x_{k-1})\|^2] \\
&\quad + 12\mathbb{E}[\|\nabla f_{j_k}(\varphi_k^{j_k}) - \nabla f_j(\xi_k^{j_k})\|^2] - 12\mathbb{E}[\|\tfrac{1}{n}\textstyle\sum_{i=1}^n \nabla f_i(\varphi_k^i) - \tfrac{1}{n}\sum_{i=1}^n \nabla f_i(\xi_k^i)\|^2] \\
&\leq \tfrac{12}{n}\sum_{i=1}^n \mathbb{E}[\|\nabla f_i(x_k) - \nabla f_i(x_{k-1})\|^2] + 12\mathbb{E}[\|\nabla f_{j_k}(\varphi_k^{j_k}) - \nabla f_{j_k}(\xi_k^{j_k})\|^2] \\
&\leq \tfrac{12}{n}\sum_{i=1}^n \mathbb{E}[\|\nabla f_i(x_k) - \nabla f_i(x_{k-1})\|^2] + 12\mathbb{E}[\|\nabla f_{j_k}(\varphi_k^{j_k}) - \nabla f_{j_k}(\xi_k^{j_k})\|^2].
\end{aligned}
$$

Equality ① is the standard variance decomposition, which states that for any random variable X, $\mathbb{E}[\|X - \mathbb{E}X\|^2] = \mathbb{E}[\|X\|^2] - \|\mathbb{E}[X]\|^2$. The second term can be reduced further by computing the expectation. The probability that $\nabla f_{j_k}(\varphi_k^{j_k}) = \nabla f_{j_{k-1}}(x_{k-1})$ is equal to the probability that $j_k = j_{k-1}$, which is $1/n$. The probability that $\nabla f_{j_k}(\varphi_k^{j_k}) = \nabla f_{j_{k-2}}(x_{k-2})$ is equal to the probability that $j_k \neq j_{k-1}$ and $j_k = j_{k-2}$, which is $\frac{1}{n}(1 - \frac{1}{n})$. Continuing in this way,

$$\begin{aligned}
&\mathbb{E}[\|\nabla f_{j_k}(\varphi_k^{j_k}) - \nabla f_{j_k}(\xi_k^{j_k})\|^2] \\
&= \tfrac{1}{n}\mathbb{E}[\|\nabla f_{j_{k-1}}(x_{k-1}) - \nabla f_{j_{k-1}}(x_{k-2})\|^2] + \tfrac{1}{n}(1 - \tfrac{1}{n})\mathbb{E}[\|\nabla f_{j_{k-2}}(x_{k-2}) - \nabla f_{j_{k-3}}(x_{k-2})\|^2] + \cdots \\
&= \tfrac{1}{n}\sum_{\ell=1}^{k-1}(1 - \tfrac{1}{n})^{k-\ell-1}\mathbb{E}[\|\nabla f_{j_\ell}(x_\ell) - \nabla f_{j_\ell}(x_{\ell-1})\|^2].
\end{aligned}$$

This implies that

$$\begin{aligned}
12\mathbb{E}[\|\nabla f_{j_k}(\varphi_k^{j_k}) - \nabla f_{j_k}(\xi_k^{j_k})\|^2] &\leq \tfrac{12}{n^2}\sum_{\ell=1}^{k-1}(1 - \tfrac{1}{n})^{k-\ell-1}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_\ell) - \nabla f_i(x_{\ell-1})\|^2] \\
&\leq \tfrac{12}{n^2}\sum_{\ell=1}^{k-1}(1 - \tfrac{1}{2n})^{k-\ell-1}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_\ell) - \nabla f_i(x_{\ell-1})\|^2].
\end{aligned}$$

We include the second inequality to simplify later arguments. This completes our bound for the first term of (19).

For the second term of (19), we recall Lemma 23.

$$\mathbb{E}[\|\widetilde{\nabla}_k^{\xi\text{-SAGA}} - \nabla f(x_{k-1})\|^2] \leq 3\sum_{\ell=1}^{k-1}(1 - \tfrac{1}{2n})^{k-\ell-1}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_\ell) - \nabla f_i(x_{\ell-1})\|^2].$$

Combining all of these bounds, we obtain

$$\begin{aligned}
\mathbb{E}[\|\widetilde{\nabla}_k^{\text{SARGE}} - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2] \leq \tfrac{12}{n}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_k) - \nabla f_i(x_{k-1})\|^2] + \tfrac{3}{2n^2}\mathbb{E}[\|\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2] \\
+ \tfrac{39}{n^2}\sum_{\ell=1}^{k-1}(1 - \tfrac{1}{2n})^{k-\ell-1}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_\ell) - \nabla f_i(x_{\ell-1})\|^2],
\end{aligned}$$

which completes the proof. □

Lemma 24 allows us to take advantage of the recursive structure of our gradient estimate. With this lemma established, we can prove a bound on the MSE.

Lemma 25 *The SARGE gradient estimator satisfies the following recursive bound:*

$$\begin{aligned}
\mathbb{E}[\|\widetilde{\nabla}_k^{\text{SARGE}} - \nabla f(x_k)\|^2] \leq (1 - \tfrac{1}{n} + \tfrac{3}{2n^2})\mathbb{E}[\|\widetilde{\nabla}_{k-1}^{\text{SARGE}} - \nabla f(x_{k-1})\|^2] + \tfrac{12}{n}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_k) - \nabla f_i(x_{k-1})\|^2] \\
+ \tfrac{39}{n^2}\sum_{\ell=1}^{k-1}(1 - \tfrac{1}{2n})^{k-\ell-1}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f_i(x_\ell) - \nabla f_i(x_{\ell-1})\|^2].
\end{aligned}$$

Proof. The beginning of our proof is similar to the proof of the variance bound for the SARAH gradient estimator in [24, Lem. 2].

$$\begin{aligned}
\mathbb{E}_k\|\widetilde{\nabla}_k^{\text{SARGE}} - \nabla f(x_k)\|^2 &= \mathbb{E}_k[\|\widetilde{\nabla}_{k-1}^{\text{SARGE}} - \nabla f(x_{k-1}) + \nabla f(x_{k-1}) - \nabla f(x_k) + \widetilde{\nabla}_k^{\text{SARGE}} - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2] \\
&= \|\widetilde{\nabla}_{k-1}^{\text{SARGE}} - \nabla f(x_{k-1})\|^2 + \|\nabla f(x_{k-1}) - \nabla f(x_k)\|^2 + \mathbb{E}_k[\|\widetilde{\nabla}_k^{\text{SARGE}} - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2] \\
&\quad + 2\langle\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}, \nabla f(x_k) - \nabla f(x_{k-1})\rangle \\
&\quad - 2\langle\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}, \mathbb{E}_k[\widetilde{\nabla}_k^{\text{SARGE}} - \widetilde{\nabla}_{k-1}^{\text{SARGE}}]\rangle \\
&\quad - 2\langle\nabla f(x_k) - \nabla f(x_{k-1}), \mathbb{E}_k[\widetilde{\nabla}_k^{\text{SARGE}} - \widetilde{\nabla}_{k-1}^{\text{SARGE}}]\rangle.
\end{aligned}$$

We consider each inner product separately. The first inner product is equal to

$$\begin{aligned}
&2\langle \nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}, \nabla f(x_k) - \nabla f(x_{k-1})\rangle \\
&= -\|\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2 - \|\nabla f(x_k) - \nabla f(x_{k-1})\|^2 + \|\nabla f(x_k) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2.
\end{aligned}$$

For the next two inner products, we use the fact that

$$\begin{aligned}
\mathbb{E}_k[\widetilde{\nabla}_k^{\text{SARGE}} - \widetilde{\nabla}_{k-1}^{\text{SARGE}}] &= \mathbb{E}_k\big[\widetilde{\nabla}_k^{\text{SAGA}} - (1-\tfrac{1}{n})\widetilde{\nabla}_k^{\xi\text{-SAGA}} + (1-\tfrac{1}{n})\widetilde{\nabla}_{k-1}^{\text{SARGE}}\big] - \widetilde{\nabla}_{k-1}^{\text{SARGE}} \\
&= \nabla f(x_k) - (1-\tfrac{1}{n})\nabla f(x_{k-1}) - \tfrac{1}{n}\widetilde{\nabla}_{k-1}^{\text{SARGE}} \\
&= \nabla f(x_k) - \nabla f(x_{k-1}) + \tfrac{1}{n}(\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}).
\end{aligned}$$

With this equality established, we see that the second inner product is equal to

$$\begin{aligned}
&-2\langle \nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}, \mathbb{E}_k[\widetilde{\nabla}_k^{\text{SARGE}} - \widetilde{\nabla}_{k-1}^{\text{SARGE}}]\rangle \\
&= -2\langle \nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}, \nabla f(x_k) - \nabla f(x_{k-1})\rangle - \tfrac{2}{n}\langle \nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}, \nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\rangle \\
&= \|\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2 + \|\nabla f(x_k) - \nabla f(x_{k-1})\|^2 \\
&\quad - \|\nabla f(x_k) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2 - \tfrac{2}{n}\|\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2 \\
&= (1-\tfrac{2}{n})\|\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2 + \|\nabla f(x_k) - \nabla f(x_{k-1})\|^2 - \|\nabla f(x_k) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2.
\end{aligned}$$

The third inner product can be bounded using a similar procedure.

$$\begin{aligned}
&-2\langle \nabla f(x_k) - \nabla f(x_{k-1}), \mathbb{E}_k[\widetilde{\nabla}_k^{\text{SARGE}} - \widetilde{\nabla}_{k-1}^{\text{SARGE}}]\rangle \\
&= -2\langle \nabla f(x_k) - \nabla f(x_{k-1}), \nabla f(x_k) - \nabla f(x_{k-1})\rangle - \tfrac{2}{n}\langle \nabla f(x_k) - \nabla f(x_{k-1}), \nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\rangle \\
&\leq -2\|\nabla f(x_k) - \nabla f(x_{k-1})\|^2 + \tfrac{1}{n}\|\nabla f(x_k) - \nabla f(x_{k-1})\|^2 + \tfrac{1}{n}\|\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2 \\
&= -(2-\tfrac{1}{n})\|\nabla f(x_k) - \nabla f(x_{k-1})\|^2 + \tfrac{1}{n}\|\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2,
\end{aligned}$$

where the inequality is Young's. Altogether and after applying the full expectation operator, we have

$$\begin{aligned}
\mathbb{E}[\|\widetilde{\nabla}_k^{\text{SARGE}} - \nabla f(x_k)\|^2] &\leq (1-\tfrac{1}{n})\mathbb{E}[\|\widetilde{\nabla}_{k-1}^{\text{SARGE}} - \nabla f(x_{k-1})\|^2] \\
&\quad - (1-\tfrac{1}{n})\mathbb{E}[\|\nabla f(x_k) - \nabla f(x_{k-1})\|^2] + \mathbb{E}[\|\widetilde{\nabla}_k^{\text{SARGE}} - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2] \\
&\leq (1-\tfrac{1}{n})\mathbb{E}[\|\widetilde{\nabla}_{k-1}^{\text{SARGE}} - \nabla f(x_{k-1})\|^2] + \mathbb{E}[\|\widetilde{\nabla}_k^{\text{SARGE}} - \widetilde{\nabla}_{k-1}^{\text{SARGE}}\|^2].
\end{aligned}$$

Finally, we bound the last term on the right using Lemma 24.

$$\begin{aligned}
\mathbb{E}\|\widetilde{\nabla}_k^{\text{SARGE}} - \nabla f(x_k)\|^2 &\leq (1-\tfrac{1}{n}+\tfrac{3}{2n^2})\mathbb{E}[\|\widetilde{\nabla}_{k-1}^{\text{SARGE}} - \nabla f(x_{k-1})\|^2] + \frac{12}{n}\sum_{i=1}^n \mathbb{E}[\|\nabla f_i(x_k) - \nabla f_i(x_{k-1})\|^2] \\
&\quad + \frac{39}{n^2}\sum_{\ell=1}^{k-1}(1-\tfrac{1}{2n})^{k-\ell-1}\sum_{i=1}^n \mathbb{E}[\|\nabla f_i(x_\ell) - \nabla f_i(x_{\ell-1})\|^2]
\end{aligned}$$

and complete the proof. □

Proof of Lemma 17 It is easy to see that $\rho_B = 1/n$ by computing the expectation of the SARGE gradient estimator.

$$\begin{aligned}
\nabla f(x_k) - \mathbb{E}_k[\widetilde{\nabla}_k^{\text{SARGE}}] &= \nabla f(x_k) - \mathbb{E}_k[\widetilde{\nabla}_k^{\text{SAGA}} - (1-\tfrac{1}{n})(\widetilde{\nabla}_k^{\xi\text{-SAGA}} - \widetilde{\nabla}_{k-1}^{\text{SARGE}})] \\
&= (1-\tfrac{1}{n})(\nabla f(x_{k-1}) - \widetilde{\nabla}_{k-1}^{\text{SARGE}}).
\end{aligned}$$

The result of Lemma 25 makes it clear that $M_1 = 12$. To determine ρ_M, we must first choose a suitable sequence $\mathcal{M}_k$. Let $\mathcal{M}_k = \mathbb{E}[\|\widetilde{\nabla}_k^{\text{SARGE}} - \nabla f(x_k)\|^2]$. If $n = 1$, then $\mathcal{M}_k = 0$ for all k, so it holds trivially that $\mathcal{M}_k \leq (1-\rho_M)\mathcal{M}_{k-1}$. If $n \geq 2$, then $1 - \frac{1}{n} + \frac{3}{2n^2} \leq 1 - \frac{1}{4n}$, so Lemma 25 ensures that with $\rho_M = \frac{1}{4n}$, $\mathcal{M}_k \leq (1-\rho_M)\mathcal{M}_{k-1}$.

Finally, we must compute M_2 and ρ_F with respect to some sequence $\mathcal{F}_k$. Lemma 25 motivates the choice

$$\mathcal{F}_k = \sum_{\ell=1}^{k-1}(1-\tfrac{1}{2n})^{k-\ell-1}\sum_{i=1}^n \mathbb{E}[\|\nabla f_i(x_\ell) - \nabla f_i(x_{\ell-1})\|^2],$$

and the choices $M_2 = \frac{39}{n}$ and $\rho_F = \frac{1}{2n}$ are clear. ■