On the Bias-Variance Tradeoff in Stochastic Gradient Methods

Derek Driggs **Jingwei Liang** **Carola-Bibiane Schönlieb**
Department of Applied Mathematics and Theoretical Physics
Cambridge University
Cambridge, England CB3 0WA
d.driggs@damtp.cam.ac.uk, {jl993,cbs31}@cam.ac.uk

Abstract

We present a general analysis of variance reduced stochastic gradient methods with bias for minimising convex, strongly convex, and non-convex composite objectives. The key to our analysis is a new connection between bias and variance in stochastic gradient estimators, suggesting a new form of bias-variance tradeoff in stochastic optimisation. This connection allows us to provide simple convergence proofs for biased algorithms, extend proximal support to biased algorithms for the first time in the convex setting, and show that biased gradient estimators often offer theoretical advantages over unbiased estimators. We propose two algorithms, B-SAGA and B-SVRG, that incorporate bias into the SAGA and SVRG gradient estimators and analyse them using our framework. Our analysis shows that the bias in the B-SAGA and B-SVRG gradient estimators decreases their mean-squared errors and improves their performance in certain settings.

1 Introduction

Consider the following convex composite minimisation problem:

$$\min_{x\in\mathbb{R}^p} \left\{ F(x) \overset{\text{def}}{=} f(x) + g(x) \overset{\text{def}}{=} \frac{1}{n}\sum\nolimits_{i=1}^{n} f_i(x) + g(x) \right\}. \tag{1}$$

We assume throughout this manuscript that f_i has an L-Lipschitz continuous gradient, g is convex, and that the operator $\text{prox}_{\eta g}(\cdot) = \arg\min_x \eta g(x) + \frac{1}{2}\|x - \cdot\|^2$ can be evaluated efficiently over the domain of g. No further restrictions on f_i or g are placed unless stated otherwise.

Problems of the form of (1) arise frequently in 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-regularised logistic regression. Principal component analysis (PCA) can also be formulated as a problem with this structure, where the functions f_i are non-convex [5, 12].

Stochastic gradient estimators In practice, the value of n is often very large, which makes classic gradient based methods, such as proximal gradient descent [7] and FISTA [6], obsolete. Recently, stochastic gradient methods have become prevalent for solving (1), since they have very low per iteration computational cost and can have the same convergence rates as their deterministic counterparts. Here, the full gradient is replaced by a stochastic gradient estimator. Let j_k be chosen randomly from $\{1, ..., n\}$. The following are popular examples of stochastic gradient estimators.

- Stochastic gradient descent [21] is the classical example, using the gradient estimator

$$\widetilde{\nabla}_{\text{SGD}} f(x_k) \overset{\text{def}}{=} \nabla f_{j_k}(x_k).$$

Preprint. Under review.

- The SAGA gradient estimator [9] has the form

$$\widetilde{\nabla}_{\text{SAGA}} f(x_k) \overset{\text{def}}{=} \nabla f_{j_k}(x_k) - \nabla f_{j_k}(\varphi_k^{j_k}) + \frac{1}{n} \textstyle\sum_{i=1}^n \nabla f_i(\varphi_k^i),$$

  where φ_k^i follows the update rule $\varphi_{k+1}^j = x_k$. The algorithms Point-SAGA [8], Finito [10], MISO [15], SDCA [23], and those in [13] use gradient estimators related to $\widetilde{\nabla}_{\text{SAGA}} f$.

- The SVRG gradient estimator [14], with s in $\{m, 2m, ...\}$, is defined as

$$\widetilde{\nabla}_{\text{SVRG}} f(x_k) \overset{\text{def}}{=} \nabla f_{j_k}(x_k) - \nabla f_{j_k}(\varphi_s) + \nabla f(\varphi_s),$$

  where φ_s is a "snapshot" point updated every m steps. The algorithms prox-SVRG [26], Katyusha [2], KatyushaX [3], Natasha [1], Natasha2 [4], MiG [27], and ASVRG [24] use the SVRG gradient estimator or a relative.

- The SARAH gradient estimator [18],

$$\widetilde{\nabla}_{\text{SARAH}} f(x_k) \overset{\text{def}}{=} \nabla f_{j_k}(x_k) - \nabla f_{j_k}(x_{k-1}) + \widetilde{\nabla}_{\text{SARAH}} f(x_{k-1}).$$

  The algorithms SARAH, prox-SARAH [19], SPIDER [11], SPIDERBoost [25] and SPIDER-M [28] use this gradient estimator.

- The SAG gradient estimator [22] is closely related to $\widetilde{\nabla}_{\text{SAGA}}$, but differs in a crucial way that we discuss below.

$$\widetilde{\nabla}_{\text{SAG}} f(x_k) \overset{\text{def}}{=} \frac{1}{n}\big(\nabla f_{j_k}(x_k) - \nabla f_{j_k}(\varphi_k^{j_k})\big) + \frac{1}{n} \textstyle\sum_{i=1}^n \nabla f_i(\varphi_k^i).$$

  There has been much less research on algorithms using this gradient estimator.

The majority of the literature on stochastic optimisation considers the first three of these examples, as these gradient estimators are *unbiased*, while the last two are *biased*.[1] 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. Biased algorithms currently have

- **Complex convergence proofs.** It is commonly believed that the relationship $\mathbb{E}_k\big[\widetilde{\nabla} f(x_k)\big] = \nabla f(x_k)$ is essential for a simple convergence analysis (see the discussion in [9], for example). The convergence proof of the unbiased algorithm SAG is notoriously complex, requiring computational verification for one of the steps [22].
- **No proximal support.** To the best of our knowledge, there are no existing theoretical guarantees for biased algorithms to solve (1) with $g \not\equiv 0$.
- **Sub-optimal convergence rates.** For example, SARAH achieves an $\mathcal{O}\big(\frac{\log(1/\epsilon)}{\epsilon}\big)$ complexity bound for solving (1) with $g \equiv 0$ and each f_i is convex [18], while SAGA/SVRG achieve a complexity bound of $\mathcal{O}\left(\frac{1}{\epsilon}\right)$.[2]

However, there are notable exceptions that suggest biased algorithms are worth further consideration. Recently, [11, 19, 25, 28] proved that algorithms using the SARAH gradient estimator achieve the oracle complexity lower bound of $\mathcal{O}\big(\frac{\sqrt{n}}{\epsilon^2}\big)$ for non-convex composite optimisation. For comparison, the best complexity proved for SAGA and SVRG in this setting is $\mathcal{O}\big(\frac{n^{2/3}}{\epsilon^2}\big)$.[3]

Contributions In this paper, we present a novel approach to study the convergence of stochastic gradient methods. The resulting framework unifies the analysis for both biased and unbiased stochastic gradient algorithms in convex, strongly convex, and non-convex settings. Our framework produces simple convergence proofs; no computational certificates are required, and it introduces a new bias-variance tradeoff in stochastic composite optimisation. We apply this framework to study

[1] The SARAH gradient estimator satisfies the relationship $\mathbb{E}\big[\widetilde{\nabla}_{\text{SARAH}} f(x_k)\big] = \mathbb{E}\left[\nabla f(x_k)\right]$, but this is not the same as being an unbiased estimator.

[2] These complexities are for finding a point satisfying $\mathbb{E}[F(x_k) - F(x^*)] \leq \epsilon$.

[3] These complexities are for finding a point satisfying $\mathbb{E}\|\mathcal{G}(x_k)\|^2 \leq \epsilon$ (see Section 3.1 for more information).

biased extensions of SAGA and SVRG, and show that incorporating bias into these gradient estimators reduces their mean squared errors which can improve their performance. Our framework allows for the development of new biased stochastic gradient estimators that can navigate the bias-variance tradeoff for better performance.

The simplicity of our unified framework comes at the cost of requiring smaller step sizes for SAG, SAGA, and SVRG than those of existing analyses. Nevertheless, we recover the state-of-the-art convergence rates for SAGA and SVRG in the convex and non-convex settings, and nearly match the state-of-the-art linear rates in the strongly convex setting.

Other related work The concurrent work [16] presents the algorithm SVAG, which is equivalent to our algorithm B-SAGA when $g \equiv 0$. The authors prove that SVAG achieves an $\mathcal{O}\left(\frac{1}{T}\right)$ convergence rate for all values of θ on convex objectives. Our analysis is simpler and more general, proving linear convergence when strong convexity is present and proving convergence rates without convexity. Our analysis also applies to many stochastic gradient estimators beyond B-SAGA, including the B-SVRG gradient estimator we consider in this work.

Notation We use $\partial g(x)$ to denote the subdifferential of g at $x \in \mathbb{R}^p$. We use $\text{prox}_{\eta g}(y)$ as shorthand for the proximal operator of ηg at $y \in \mathbb{R}^p$; we provide a definition of the proximal operator in Appendix B. For a general stochastic gradient operator, we use $\widetilde{\nabla}$, and we include subscripts to refer to specific estimators (e.g. $\widetilde{\nabla}_{\text{B-SAGA}}$). The points φ_k^i denote points where an algorithm has evaluated a stored gradient of f_i. The operator $\mathbb{E}_k$ is expectation conditioned on the random variables $j_1, j_2, \cdots, j_{k-1}$.

2 Bias and Variance in Stochastic Gradient Estimation

In this section, we outline the role of bias in stochastic gradient methods. Bias affects the convergence analysis of stochastic gradient methods in two respects. The first is its affect on the *mean squared error (MSE)* of the gradient estimator.

Definition 1. *The* mean squared error *of the stochastic gradient estimator* $\widetilde{\nabla} f(x_k)$ *is defined as*

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

The MSE admits the following bias-variance decomposition*:*

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

The decomposition in (1) shows that introducing bias to the estimator can decrease the MSE if the biased estimator has a smaller variance; we refer to a relationship of this type as a *bias-variance tradeoff.*

Bias also manifests as an additional term in our convergence analysis, which we refer to as the "bias term". Below, we outline how the MSE and bias term arise from the traditional analysis of gradient descent methods for non-composite, composite, and non-convex objectives, as the effects differ slightly in each case.

Non-composite case ($g \equiv 0$) Consider applying the vanilla gradient descent to solve (1) with $g \equiv 0$ and f_i convex. Let $\eta \leq 2/L$ be the step size. The Lipschitz continuity of ∇f implies [17],

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

which is the classic descent property of gradient descent. When using a stochastic gradient estimator, we can only obtain the following relation:

$$\begin{aligned}
&\mathbb{E}_k\left[f(x_{k+1}) - f(x^*)\right] = \mathbb{E}_k\left[f(x_{k+1}) - f(x_k) + f(x_k) - f(x^*)\right] \\
&\quad \leq \tfrac{1}{\eta}\mathbb{E}_k\langle \nabla f(x_k) - \widetilde{\nabla} f(x_k), x_{k+1} - x_k\rangle + \left(\tfrac{L}{2} - \tfrac{1}{\eta}\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2 + f(x_k) - f(x^*) \\
&\quad \leq \tfrac{\epsilon}{2\eta}\mathbb{E}_k\|\nabla f(x_k) - \widetilde{\nabla} f(x_k)\|^2 + \left(\tfrac{L}{2} + \tfrac{1}{2\eta\epsilon} - \tfrac{1}{\eta}\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2 + f(x_k) - f(x^*). \quad (3)
\end{aligned}$$

Compared to gradient descent, the difficulty of analysing stochastic gradient methods is that we must control the MSE of the gradient estimator. In the non-composite case, it has been suggested that biased estimators such as SAG might perform better than unbiased estimators because they have smaller variance (see, for example, the discussion in [9]). While the SAG estimator does have smaller variance, it is the MSE that arises in our analysis, so we should not necessarily infer a relationship between the variance of a gradient estimator and its performance. Instead, we must minimise the variance and the bias of our gradient estimator in order to minimise the MSE through the relationship (1).

Composite case ($g \not\equiv 0$) The situation becomes more complicated when g is non-trivial. Let $G_{k+1} \in \partial g(x_{k+1})$ be a subgradient, then for proximal stochastic gradient descent, we have

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

Inequality ① is an application of (2), ② follows from the convexity of g, and ③ comes from the implicit definition of the proximal operator (see (B)). The term $\langle \nabla f(x_k) - \widetilde{\nabla} f(x_k), x_k - x^*\rangle$ vanishes when $\widetilde{\nabla} f(x_k)$ is an unbiased estimator, which is why unbiased algorithms are easier to analyse. When the estimator is biased, we must develop a new way to control this term.

At this point, we have uncovered a bias-variance tradeoff in stochastic composite optimisation that subsumes the tradeoff in (1). We have two terms unique to the analysis of stochastic methods,

$$
\mathbb{E}_k\langle \nabla f(x_k) - \widetilde{\nabla} f(x_k), x_k - x^*\rangle \quad \text{and} \quad \mathbb{E}_k\|\nabla f(x_k) - \widetilde{\nabla} f(x_k)\|^2, \tag{5}
$$

that must be bounded to ensure convergence. We refer to the first term as the "bias term", and the second term is simply the MSE defined above. The optimal stochastic gradient estimator minimises the effects of both of these terms. This relationship, together with (1), forms a new bias-variance tradeoff in stochastic composite optimisation.

Non-convex case The influence of bias is much simpler in non-convex composite optimisation, which explains why biased algorithms have recently had success in this regime. In the non-convex case, only the later of the two terms in (2) appears in our convergence analysis. The optimal bias must only minimise the MSE, so navigating the bias-variance tradeoff amounts to adjusting the bias and variance of $\widetilde{\nabla} f(x_k)$ to affect the MSE through the relationship in (1).

Analysis framework The challenge in analysing the convergence of biased gradient schemes is to find biased gradient estimators that minimise the effects of these two terms in (2) simultaneously. This work takes a step in this direction. Based on the discussion above, our proposed simple framework for analysing biased stochastic gradient methods is summarised below:

A framework for analysing biased stochastic gradient methods for convex composite objectives

1. Apply inequality (2) to bound the expected suboptimality $\mathbb{E}_k\left[F(x_{k+1}) - F(x^*)\right]$.
2. Derive a bound on the MSE involving $\|x_{k+1} - x_k\|^2$ and telescoping terms.
3. Derive a bound on the bias term involving $\|x_{k+1} - x_k\|^2$ and telescoping terms.
4. Sum the resulting inequality from $k = 0$ to $k = T - 1$, obtaining a bound on the suboptimality of the average iterate $\overline{x} \overset{\text{def}}{=} \frac{1}{T}\sum_{k=1}^{T} x_k$. This provides a convergence rate of $\mathcal{O}\left(\frac{1}{T}\right)$.

When the objective function is strongly convex, a slight modification to step four proves a linear convergence rate. In the non-convex setting, a similar process produces a convergence rate to a first-order stationary point. The bulk of our contribution is in steps two and three, where we provide bounds on the bias and the MSE for several stochastic gradient estimators that are compatible with this four-step framework. Our bounds for the bias term are particularly useful; these bounds allow us to extend proximal support to biased algorithms.

3 Two Biased Extensions of SAGA and SVRG

To demonstrate the usefulness of our framework, we present two biased extensions of SAGA and SVRG algorithms and derive convergence rates using the framework above. To extend the gradient estimators of SAGA and SVRG to the biased setting, we propose the following generalisation: let $\theta > 0$ be positive, then for SAGA we consider the gradient estimator

$$\widetilde{\nabla}_{\text{B-SAGA}} f(x_k) \overset{\text{def}}{=} \frac{1}{\theta}\big(\nabla f_j(x_k) - \nabla f_j(\varphi_k^j)\big) + \frac{1}{n}\sum_{i=1}^n \nabla f_i(\varphi_k^i),$$

and for SVRG we consider

$$\widetilde{\nabla}_{\text{B-SVRG}} f(x_k) \overset{\text{def}}{=} \frac{1}{\theta}\left(\nabla f_j(x_k) - \nabla f_j(\varphi_s)\right) + \nabla f(\varphi_s).$$

The algorithms resulting from the above two biased gradient estimators are described in Algorithms 1 and 2.

Algorithm 1 B-SAGA

```
Input: Step size $\eta$ and bias parameter $\theta$ set as in Theorem 3 or Theorem 5.
Output: If $F$ is convex, output $x_T$. Otherwise, output $x_\alpha$ with $\alpha$ chosen uniformly at random
    from the set $\{1, 2, \cdots, T\}$.
Initialise $x_0$ to a random value and compute $\nabla f_i(\varphi_0^i)$ for $\varphi_0^i = x_0$.
for $k = 0, 1, \cdots, T-1$ do
    Choose index $j_k \in \{1, 2, \cdots, n\}$ uniformly at random.
    $\widetilde{\nabla}_{\text{B-SAGA}} f(x_k) \leftarrow \frac{1}{\theta}\big(\nabla f_{j_k}(x_k) - \nabla f_{j_k}(\varphi_k^{j_k})\big) + \frac{1}{n}\sum_{i=1}^n \nabla f_i(\varphi_k^i)$
    $x_{k+1} \leftarrow \text{prox}_{\eta g}\big(x_k - \eta \widetilde{\nabla}_{\text{B-SAGA}} f(x_k)\big)$.
    Replace the stored gradient $\nabla f_{j_k}(\varphi_k^{j_k})$ with $\nabla f_{j_k}(x_k)$.
end for
```

Algorithm 2 B-SVRG

```
Input: Step size $\eta$ and bias parameter $\theta$ set as in Theorem 4 or Theorem 6; epoch size $m$.
Output: If $F$ is convex, output $x_{mS}$. Otherwise, output $x_\alpha$ with $\alpha$ chosen uniformly at random
    from the set $\{1, 2, \cdots, mS\}$.
Initialise $x_0$ to a random value.
$\varphi_0 \leftarrow x_0$.
for $s = 0, 1, \cdots, S-1$ do
    $\nabla f(\varphi_s) \leftarrow \nabla f(x_k)$.
    for $k = ms+1, ms+2, \cdots, m(s+1)$ do
        Choose index $j_k \in \{1, 2, \cdots, n\}$ uniformly at random.
        $\widetilde{\nabla}_{\text{B-SVRG}} f(x_k) \leftarrow \frac{1}{\theta}\big(\nabla f_{j_k}(x_k) - \nabla f_{j_k}(\varphi_s)\big) + \nabla f(\varphi_s)$
        $x_{k+1} \leftarrow \text{prox}_{\eta g}\big(x_k - \eta \widetilde{\nabla}_{\text{B-SVRG}} f(x_k)\big)$.
    end for
end for
```

Remark 1. *The B-SAGA and B-SVRG gradient estimators differ from those in SAGA and SVRG by giving more weight to stored gradients from previous iterations. The amount of weight is determined through the parameter $\theta > 0$. As θ increases, the B-SAGA gradient estimator increasingly favors the average of the stored gradients $\frac{1}{n}\sum_{i=1}^n \nabla f_i(\varphi_k^i)$, and the B-SVRG gradient estimator increasingly favors the full gradient of the "snapshot" point φ_s. For $\theta = n$, the B-SAGA gradient estimator is equivalent to the SAG gradient estimator.*

3.1 Convergence Rates for B-SAGA and B-SVRG

In this section, we discuss the convergence rates of B-SAGA and B-SVRG for the cases that problem (1) is convex, strongly convex and non-convex.

- When the objective function is convex, we prove convergence rates with respect to the suboptimality $\mathbb{E}[F(x_{k+1}) - F(x^*)] \leq \epsilon$.
- When the objective function is strongly convex, we prove convergence rates with respect to the distance from the optimiser $\mathbb{E}\|x_k - x^*\|^2$.
- For non-convex objectives, we measure convergence to a first-order stationary point defined with respect to the *generalised gradient map* $\mathcal{G}(x_k) \overset{\text{def}}{=} \frac{1}{\eta}\left(x_k - \operatorname{prox}_{\eta g}\left(x_k - \eta\nabla f(x_k)\right)\right)$. Our measure of convergence is the norm of the generalised gradient $\mathbb{E}\|\mathcal{G}(x_k)\|^2$.

For B-SAGA and B-SVRG, when $\theta \neq 1$, both gradient estimators are biased, and the amount of bias is reflected in the parameter θ. Consider the B-SAGA gradient estimator as an example.

$$\mathbb{E}_k\left[\widetilde{\nabla}_{\text{B-SAGA}} f(x_k)\right] - \nabla f(x_k) = \left(1 - \tfrac{1}{\theta}\right)\left(\textstyle\sum_{i=1}^{n} \nabla f_i(\varphi_k^i) - \nabla f(x_k)\right).$$

The same equality holds for the B-SVRG gradient estimator, recognising that $\varphi_k^i = \varphi_s$. Our convergence analysis relies on finding useful bounds for the two terms in equation (2). The following lemma provides a bound for the bias term.

Lemma 1. *Suppose g is μ-strongly convex with $\mu \geq 0$, and set the bias parameter $\theta \geq 1$. Let $\lambda > 0$ be a constant whose value we determine later and the operator $\widetilde{\nabla} \equiv \widetilde{\nabla}_{\text{B-SAGA}}$ or $\widetilde{\nabla}_{\text{B-SVRG}}$. The following inequality holds:*

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

When $\widetilde{\nabla} \equiv \widetilde{\nabla}_{\text{B-SVRG}}$, we can replace each φ_k^i with φ_s for k in epoch s.

The inequality of Lemma 1 is the same inequality as the one in Section 2, but we have replaced the bias term

$$\mathbb{E}_k\langle \nabla f(x_k) - \widetilde{\nabla} f(x_k), x_k - x^*\rangle \quad \text{with} \quad \left(1 - \tfrac{1}{\theta}\right)\textstyle\sum_{i=1}^{n}\|x_k - \varphi_k^i\|^2.$$

The next lemma bounds the MSE.

Lemma 2. *Let the operator $\widetilde{\nabla} \equiv \widetilde{\nabla}_{\text{B-SAGA}}$ or $\widetilde{\nabla}_{\text{B-SVRG}}$. The MSE of these stochastic gradient estimators satisfy*

$$\begin{aligned}&\mathbb{E}_k\|\widetilde{\nabla} f(x_k) - \nabla f(x_k)\|^2 \\ &\leq \tfrac{L^2}{n\theta^2}\textstyle\sum_{i=1}^{n}\|x_k - \varphi_k^i\|^2 + \left(1 - \dfrac{2}{\theta}\right)\|\nabla f(x_k) - \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\nabla f_i(\varphi_k^i)\|^2.\end{aligned}$$

When $\widetilde{\nabla} \equiv \widetilde{\nabla}_{\text{B-SVRG}}$, we can replace each φ_k^i with φ_s for k in epoch s.

Convex objectives After applying Lemma 1 to bound the bias term and Lemma 2 to bound the MSE, we apply algorithm-specific telescoping procedures to prove convergence guarantees for B-SAGA and B-SVRG in the convex and strongly convex settings.

Theorem 3 (B-SAGA). *In Algorithm 1, set $\eta = \frac{\theta}{4Ln^2(\theta-1)+L(\theta+3\sqrt{2}n)}$ for $\theta \in [1,2)$, and set $\eta = \frac{\theta}{4Ln^2(\theta-1)+3\sqrt{2}Ln(\theta-1)+L\theta}$ for $\theta \geq 2$. After T iterations of Algorithm 1, we have the following bound on the suboptimality of the average iterate $\overline{x} \overset{def}{=} \frac{1}{T}\sum_{k=1}^{T} x_k$:*

$$\mathbb{E}\left[F(\overline{x}) - F(x^*)\right] \leq \begin{cases} \frac{2Ln^2(\theta-1)+\frac{L}{2}(\theta+3\sqrt{2}n)}{\theta T}\|x_0 - x^*\|^2 & \theta \in [1,2), \\ \frac{2Ln^2(\theta-1)+\frac{3\sqrt{2}}{2}Ln(\theta-1)+L\theta}{\theta T}\|x_0 - x^*\|^2 & \theta \geq 2. \end{cases}$$

If g is μ-strongly convex, after T iterations, Algorithm 1 produces an iterate satisfying

$$\mathbb{E}\|x_T - x^*\|^2 \leq \begin{cases} \left(1 + \frac{\mu\theta}{8Ln^2(\theta-1)+2L(\theta+3\sqrt{2}n)}\right)^{-T} \|x_0 - x^*\|^2 & \text{for } \theta \in [1,2) \\ \left(1 + \frac{\mu\theta}{8Ln^2(\theta-1)+6\sqrt{2}Ln(\theta-1)+2L\theta}\right)^{-T} \|x_0 - x^*\|^2 & \text{for } \theta \geq 2. \end{cases}$$

Theorem 4 (B-SVRG). *In Algorithm 2, set $\eta = \frac{\theta}{4Lm(m+1)(\theta-1)+L(\theta+3\sqrt{2m(m+1)}}$ for $\theta \in [1,2)$ and set $\eta = \frac{\theta}{4Lm(m+1)(\theta-1)+3L\sqrt{2m(m+1)}(\theta-1)+L\theta}$ for $\theta \geq 2$. After S epochs of Algorithm 2, we have the following bound on the suboptimality of the average iterate:*

$$\mathbb{E}\left[F(\overline{x}) - F(x^*)\right] \leq \begin{cases} \frac{2Lm(m+1)(\theta-1)+\frac{L}{2}(\theta+3\sqrt{2m(m+1)})}{mS\theta} \|x_0 - x^*\|^2 & \theta \in [1,2), \\ \frac{2Lm(m+1)(\theta-1)+\frac{3\sqrt{2}}{2}L\sqrt{m(m+1)}(\theta-1)+\frac{L\theta}{2}}{mS\theta} \|x_0 - x^*\|^2 & \theta \geq 2. \end{cases}$$

If g is μ-strongly convex, set $\eta = \frac{\theta}{5Lm(m+1)(\theta-1)+\frac{5L}{4}(\theta+3\sqrt{2m(m+1)})}$ for $\theta \in [1,2)$, and set $\eta = \frac{\theta}{5Lm(m+1)(\theta-1)+L\sqrt{2m(m+1)}(\theta-1)+L\theta}$. After S epochs, Algorithm 2 produces an iterate satisfying

$$\mathbb{E}\|x_{mS} - x^*\|^2 \leq \begin{cases} \left(1 + \frac{\mu\theta}{10Lm(m+1)(\theta-1)+\frac{5L}{2}(\theta+3\sqrt{2m(m+1)})}\right)^{-mS} \|x_0 - x^*\|^2 & \text{for } \theta \in [1,2) \\ \left(1 + \frac{\mu\theta}{10Lm(m+1)(\theta-1)+2L\sqrt{2m(m+1)}(\theta-1)+2L\theta}\right)^{-mS} \|x_0 - x^*\|^2 & \text{for } \theta \geq 2. \end{cases}$$

This analysis reveals the benefits of bias on the MSE of these gradient estimators. The value of θ that minimises the MSE depends on the relative values of $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_k) - \nabla f_i(\varphi_k^i)\|^2$ and $\|\nabla f(x_k) - \frac{1}{n}\sum_{i=1}^n \nabla f_i(\varphi_k^i)\|^2$, but our analysis suggests that setting $\theta \approx 2$ roughly minimises the MSE.

However, the bias term of equation (2) counteracts the smaller MSE. While the influence of the variance decreases with η, the influence of the bias stays constant. This causes the bias term of equation (2) to outweigh the benefits of a smaller MSE. Because our analysis requires smaller step sizes than existing analyses for SAGA, SAG, and SVRG, it could be that these theoretical benefits of bias are pessimistic as well.

Remark 2. *Although we consider only the B-SAGA and B-SVRG gradient estimators in this work, our analysis extends to many gradient estimators, including those presented in [13], with only small changes in the MSE and bias term bounds. This allows many algorithms to incorporate bias into their gradient estimators.*

Non-convex objectives The theoretical benefits of bias are even more pronounced in the non-convex regime. For B-SAGA and B-SVRG in the non-convex setting, we prove the following guarantees for convergence to a first-order stationary point.

Theorem 5 (B-SAGA). *In Algorithm 1, set $\eta = \frac{\theta}{2Ln}$ for $\theta \leq 2$, and set $\eta = \frac{\theta}{2Ln(\theta-1)}$ for $\theta \geq 2$. After T steps, Algorithm 1 achieves the following bound on the norm of the generalised gradient:*

$$\mathbb{E}\|\mathcal{G}_\eta(x_\alpha)\|^2 \leq \begin{cases} \frac{4Ln(F(x_0)-F(x^*))}{\theta\left(1-\frac{\theta}{n}\right)T} & \text{for } 0 < \theta < 2, \\ \frac{4Ln(\theta-1)(F(x_0)-F(x^*))}{\theta\left(1-\frac{\theta}{n(\theta-1)}\right)T} & \text{for } \theta \geq 2. \end{cases}$$

Theorem 6 (B-SVRG). *In Algorithm 2, set $\eta = \frac{\theta}{2L\sqrt{m(m+1)}}$ for $\theta \leq 2$, and set $\eta = \frac{\theta}{2Ln(\theta-1)\sqrt{m(m+1)}}$ for $\theta \geq 2$. After T steps, Algorithm 2 achieves the following bound on the norm of the generalised gradient:*

$$\mathbb{E}\|\mathcal{G}_\eta(x_\alpha)\|^2 \leq \begin{cases} \frac{4L\sqrt{m(m+1)}(F(x_0)-F(x^*))}{\theta\left(1-\frac{\theta}{\sqrt{m(m+1)}}\right)T} & \text{for } 0 < \theta < 2, \\ \frac{4L\sqrt{m(m+1)}(\theta-1)(F(x_0)-F(x^*))}{\theta\left(1-\frac{\theta}{(\theta-1)\sqrt{m(m+1)}}\right)T} & \text{for } \theta \geq 2. \end{cases}$$

These results reflect the difference in the effect that bias has in convex and non-convex optimisation, just as we discussed in Section 2. Our analysis suggests that the convergence rates of B-SAGA and B-SVRG in the non-convex setting are optimised for $\theta = 2$, when the MSE is roughly minimised. This differs from the convex setting because the bias term of (2) no longer affects the convergence analysis. Theorems 5 and 6 also prove convergence rates for convex problems when $\theta < 1$, a regime that Theorems 3 and 4 do not cover.

4 Numerical Experiments

In this section, we present numerical experiments testing B-SAGA and B-SVRG on convex and strongly convex objectives. We include experiments for non-convex objectives in Appendix A. To test the influence of bias in the B-SAGA and B-SVRG gradient estimators, we used these algorithms to solve a series of ridge regression and LASSO tasks. Let $(h_i, l_i) \in \mathbb{R}^p \times \{\pm 1\},\ i = 1, \cdots, n$ be the training set, where $l_i \in \mathbb{R}^p$ is the feature vector of each data sample, and l_i is the binary label. Let $\alpha > 0$ be a tuning parameter. The ridge regression problem takes the form

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

LASSO is similar, but with the regulariser $\|x\|_1$ replacing $\|x\|_2^2$. These problems are of the form (1), where we set g equal to the regulariser. In ridge regression, g is strongly convex, and in LASSO, g is convex but not strongly convex.

We consider four binary classification data sets: `australian`, `mushrooms`, `phishing`, and `ijcnn1` from LIBSVM. [4] In all our experiments, we rescale the value of the data to $[-1, 1]$, set $\alpha = \frac{1}{n}$, and tune the step size to achieve the best performance for the algorithm with $\theta = 1$ (the unbiased case).

We consider $\theta \in \{1, 10, 100, n\}$ for B-SAGA and $\theta \in \{0.5, 0.8, 1, 1.5\}$ for B-SVRG and measure performance with respect to the suboptimality $F(x_k) - F(x^*)$, where x^* is a low-tolerance solution found using proximal gradient descent. For B-SVRG, we use the epoch length $m = 2n$. We perform all experiments using MATLAB R2019a on a machine with four cores, 25 GB of RAM, and a clock speed of 3.40 GHz. Plots of these experiments are included in Appendix A. Our experiments suggest that

- B-SAGA consistently performs better with moderate amounts of bias (i.e. $\theta \in (1, n)$).
- For a fixed step size, B-SVRG is much more sensitive to θ than B-SAGA. Small amounts of bias (i.e. $\theta \in [0.8, 1.5]$) can occasionally improve performance. The benefits of bias in B-SVRG are more apparent in the non-convex setting, shown in Appendix A.

The above observations indicate that, depending on the data, biased schemes can benefit from their biased gradient estimates, as the free parameter θ reduces the MSE of the gradient estimators leading to better performance.

5 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 minimising 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 through the traditional bias-variance decomposition of MSE. However, we show that the convergence rates of biased algorithms depend on a new bias-variance tradeoff that subsumes the tradeoff in the MSE alone. Future work can use the framework presented in this manuscript to develop new biased gradient estimators that navigate this bias-variance tradeoff for improved performance and end our dependence on unbiased estimators.

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

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 Yuan, Y. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. In *ICML* (2018).

[6] 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.

[7] Combettes, P. L., and Pesquet, J.-C. Proximal splitting methods in signal processing. *Inverse Problems in Science and Engineering* (2012), 185–212.

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

[9] 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.

[10] 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).

[11] 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).

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

[13] 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.

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

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

[16] Morin, M., and Giselsson, P. SVAG: Unified convergence results for sag-saga interpolation with stochastic variance adjusted gradient descent. *arXiv:1903.09009* (2019).

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

[18] 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.

[19] 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).

[20] Reddi, S. J., Sra, S., Póczós, B., and Smola, A. Fast stochastic methods for nonsmooth nonconvex optimization. In *ICML* (2016).

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

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

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

[24] 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.

[25] 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).

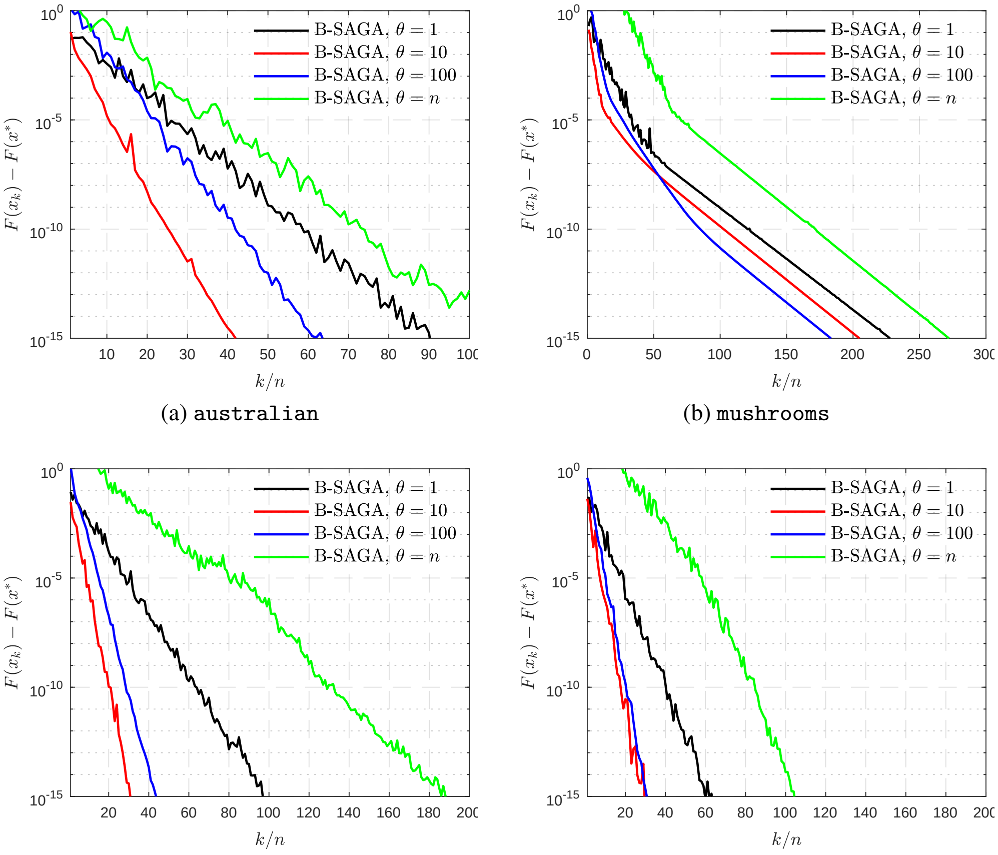
[26] Xiao, L., and Zhang, T. A proximal stochastic gradient method with progressive variance reduction. *Technical report, Microsoft Research* (2014).

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

[28] 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).

A Further Numerical Experiments

A.1 B-SAGA: Ridge Regression



(a) `australian`

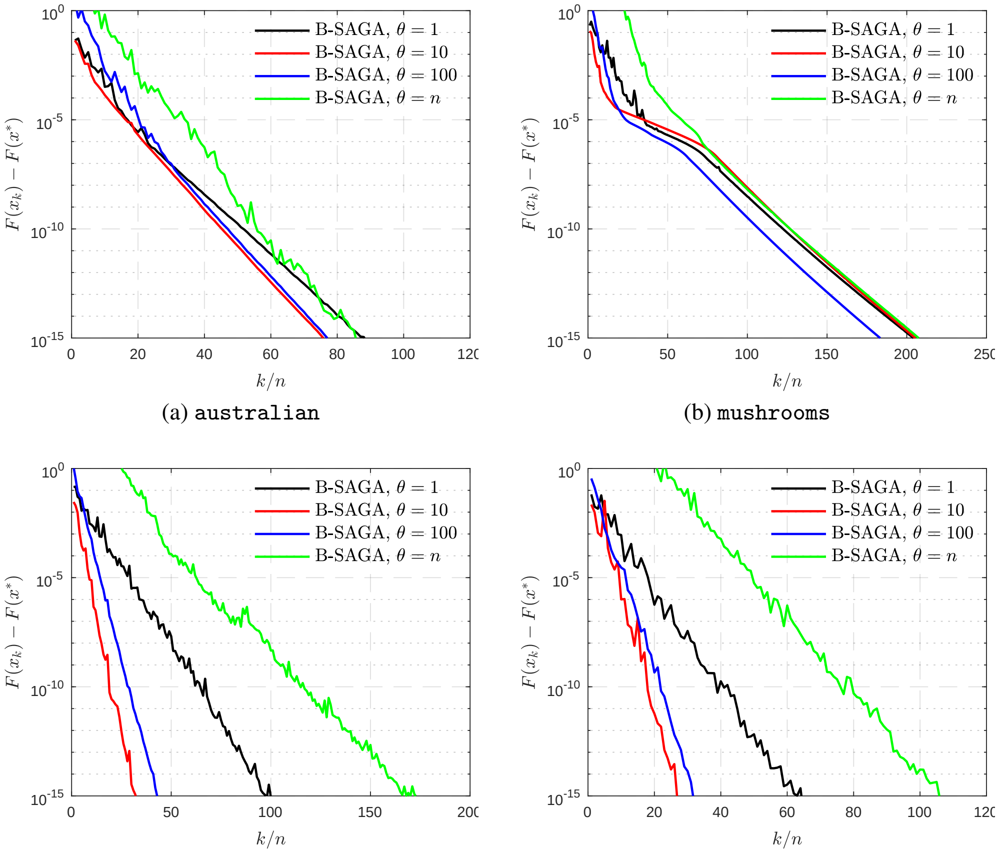
(b) `mushrooms`

(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.2 B-SAGA: LASSO



(a) `australian`

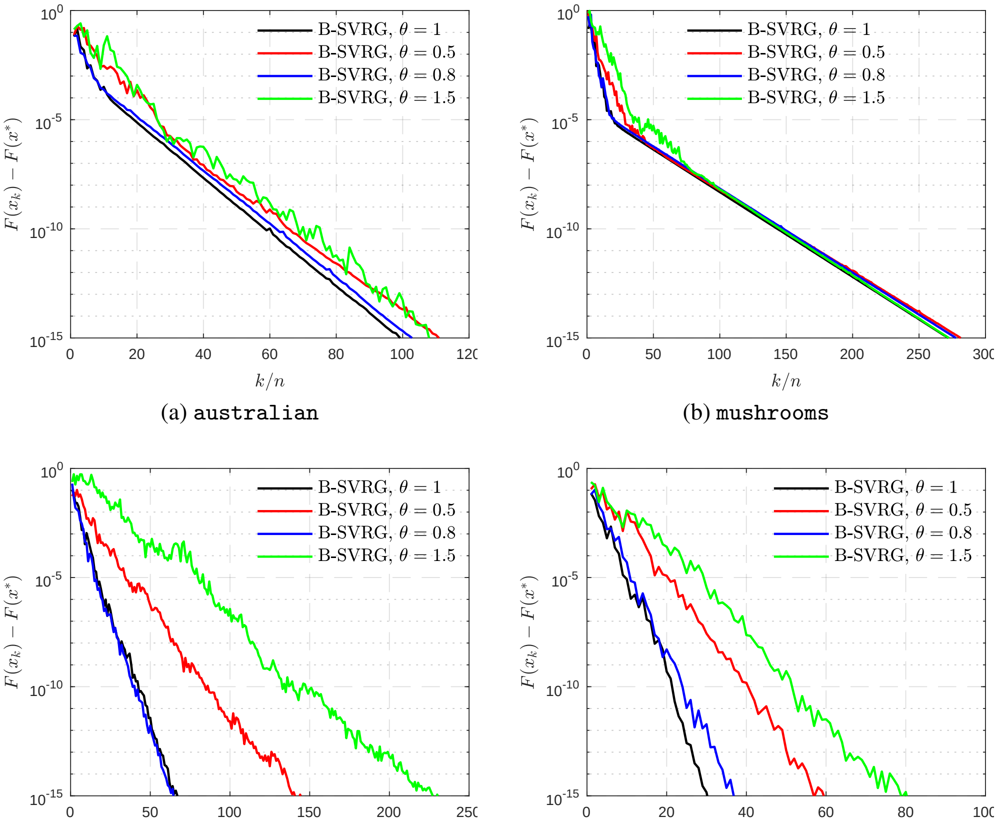
(b) `mushrooms`

(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.3 B-SVRG: Ridge Regression



(a) `australian`

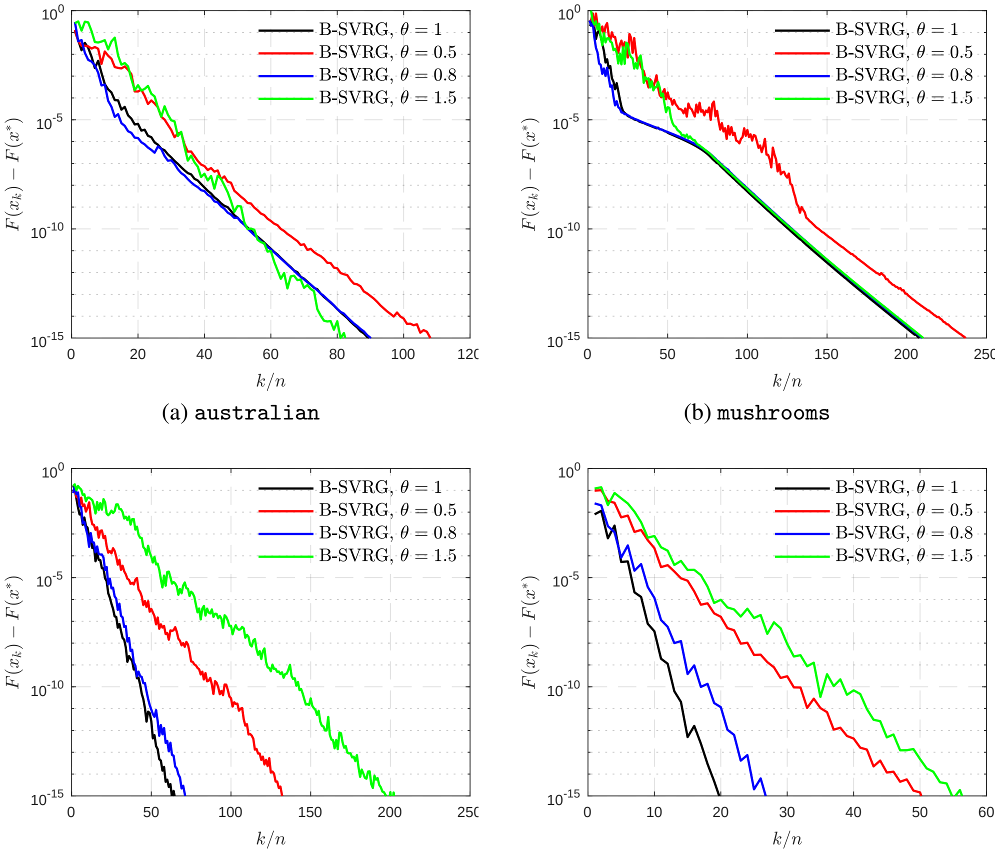
(b) `mushrooms`

(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.4 B-SVRG: LASSO



(a) australian

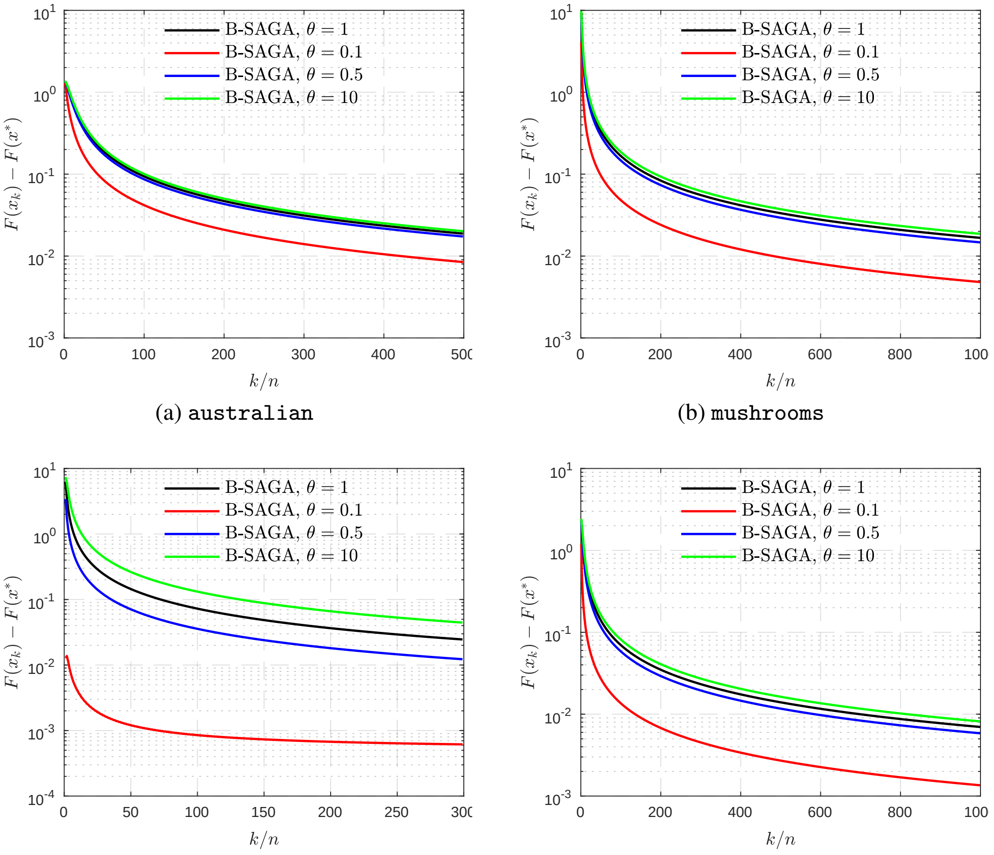
(b) mushrooms

(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 $\eta = \frac{1}{5L}$.

A.5 Experiments on non-convex objectives



(a) `australian`

(b) `mushrooms`

(c) `phishing`

(d) `ijcnn1`

Figure 5: 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^* is found by solving the problem using proximal gradient descent.

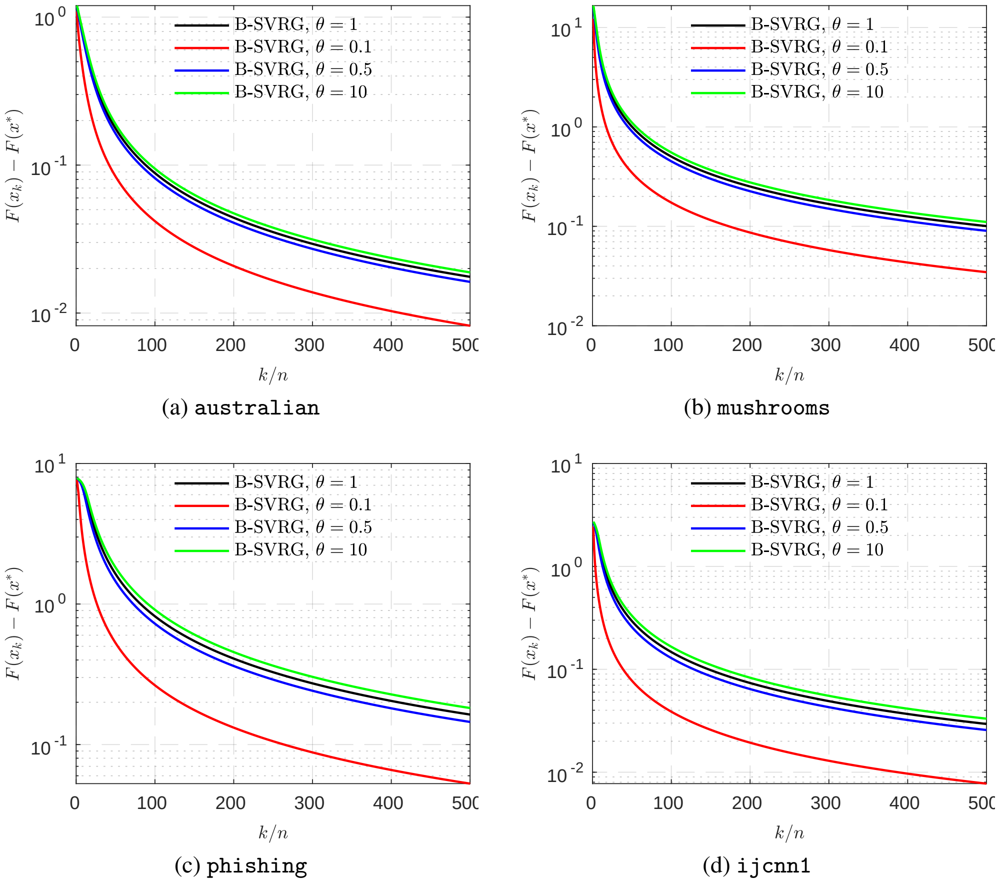
To test the effect of bias in the non-convex setting, we apply B-SAGA and B-SVRG to solve a series of non-negative principal component analysis (NN-PCA) problems. We formulate NN-PCA as in [20].

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

Here, we use $\iota_C(x)$ to denote the indicator function of the set C, and we define C to be $\{x \in \mathbb{R}^p : \|x\| \leq 1,\ x \geq 0\}$, the intersection of the unit ball and the non-negative cone. This problem is of the form (1). Letting $g \equiv \iota_C$, the operator $\text{prox}_{\eta g}$ is the projection onto C, which can be computed efficiently.

Because this problem is non-convex, we cannot measure convergence with respect to the global optimum x^*, so we use many iterations of proximal gradient descent with a small step size ($\eta = \frac{1}{10Ln}$) to find a reference point x^*. Every test is initialised 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. For B-SVRG, we use the epoch length $m = 2n$. We report the suboptimality $F(x_k) - F(x^*)$ averaged over every n iterations. These experiments suggest the following trends:

- The performance of B-SAGA varies significantly with θ, with smaller values leading to better performance.



(a) `australian`

(b) `mushrooms`

(c) `phishing`

(d) `ijcnn1`

Figure 6: 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^* is found by solving the problem using proximal gradient descent.

- B-SVRG also improves with smaller values of θ, but the improvement is often less dramatic than it is for B-SAGA in the first few epochs.

For both B-SAGA and B-SVRG, there are clear benefits to using biased gradient estimates.

B More on the proximal operator

In Section 2, we use the implicit definition of the proximal operator in our consideration of the case $g \not\equiv 0$. We provide this definition here.

Definition 2. *The proximal operator* $\text{prox}_{\eta g} : \mathbb{R}^p \to \text{dom}\, \partial g$ *is defined as*

$$\text{prox}_{\eta g}(y) \overset{def}{=} \underset{x \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2\eta} \|x - y\|^2 + g(x) \right\}.$$

Equivalently, the proximal operator is defined implicitly as the unique map satisfying

$$y - \text{prox}_{\eta g}(y) \in \partial g(\text{prox}_{\eta g}(y)).$$

Combining this implicit definition of the proximal operator with the definition of x_{k+1} in Algorithms 1 and 2, we have

$$x_k - x_{k+1} - \eta \widetilde{\nabla} f(x_k) \in \partial g(x_{k+1}), \tag{6}$$

where $\widetilde{\nabla} \equiv \widetilde{\nabla}_{\text{B-SAGA}}$ or $\widetilde{\nabla}_{\text{B-SVRG}}$.

C Elementary Lemmas

These lemmas are basic results in convex analysis that we provide for completeness.

Lemma 7. *Suppose f is convex with an L-Lipschitz continuous gradient. We have for all $x, u \in \mathbb{R}^p$,*

$$\|\nabla f(x) - \nabla f(u)\|^2 \leq 2L \left[f(x) - f(u) - \langle \nabla f(u), x - u \rangle \right]$$

Proof. We refer to [17, Thm 2.1.5] for a proof. □

Lemma 8. *Let $f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$, where each f_i is convex with an L-Lipschitz continuous gradient. Then for every $x, u \in \mathbb{R}^p$,*

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

Proof. This follows immediately from applying Lemma 7 to each component f_i. □

Lemma 9. *Suppose vectors $x, y, d \in \mathbb{R}^p$ satisfy*

$$y = \mathrm{prox}_{\eta g} \left(x - \eta d \right).$$

Then for all z, the following inequality holds:

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

Proof. We refer to [20, Lem. 2] for a proof. □

D Proof of Lemma 1

In this section, we prove the general inequality of Lemma 1 that is fundamental to our analysis of B-SAGA and B-SVRG. This inequality holds for many other gradient estimators as well, including those presented in [13], allowing bias to be incorporated into many stochastic gradient algorithms.

This first lemma is a standard result on proximal mirror descent.

Lemma 10 (Mirror Descent). *Suppose g is μ-strongly convex. Let*

$$x_{k+1} = \arg\min_y \left\{ \frac{1}{2\eta} \|y - x_k\|^2 + \langle \widetilde{\nabla} f(x_k), y - x_k \rangle - g(y) \right\}.$$

With $\widetilde{\nabla} f(x_k)$ fixed and for any $u \in \mathbb{R}^p$, we have

$$\begin{aligned} \eta \langle \widetilde{\nabla} f(x_k), x_{k+1} - u \rangle \leq & \frac{1}{2} \|x_k - u\|^2 - \frac{1 + \mu\eta}{2} \|x_{k+1} - u\|^2 - \frac{1}{2} \|x_{k+1} - x_k\|^2 \\ & - g(x_{k+1}) + g(u) \end{aligned}$$

Proof. We refer to [2], Lemma 3.5 for a proof. □

The next lemma follows the traditional analysis of gradient descent. It provides a lower bound on the amount of progress that gradient descent makes in a single iteration. Because we are using a stochastic estimate of the gradient, this lower bound includes the MSE of our gradient estimator. Unlike gradient descent, stochastic gradient methods are not guaranteed to decrease the suboptimality at each iteration—even in expectation—unless the MSE of the gradient estimator can be controlled.

Lemma 11 (Gradient Descent). *Let x_{k+1} be as defined in Lemma 10, and let $\lambda > 0$ be a constant whose value we determine later. Define*

$$\mathrm{Prog}(x_k) \overset{def}{=} - \left(\frac{1}{2\eta} \|x_{k+1} - x_k\|^2 + \langle \widetilde{\nabla} f(x_k), x_{k+1} - x_k \rangle + g(x_{k+1}) \right).$$

Then

$$\begin{aligned}\mathbb{E}_k[\text{Prog}(x_k)] \leq & f(x_k) - \mathbb{E}_k[F(x_{k+1})] + \frac{1}{2L\lambda}\mathbb{E}_k\|\nabla f(x_k) - \widetilde{\nabla} f(x_k)\|^2 \\ & + \left(\frac{L(\lambda+1)}{2} - \frac{1}{2\eta}\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2\end{aligned}$$

Proof. We begin with the definition of $\text{Prog}(x_k)$.

$$\begin{aligned}& - \mathbb{E}_k\left[\frac{1}{2\eta}\|x_{k+1} - x_k\|^2 + \langle \widetilde{\nabla} f(x_k), x_{k+1} - x_k\rangle + g(x_{k+1})\right] \\ &= \mathbb{E}_k\Bigg[- \left(\frac{L}{2}\|x_{k+1} - x_k\|^2 + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + g(x_{k+1})\right) \\ &\qquad + \left(\left\langle \nabla f(x_k) - \widetilde{\nabla} f(x_k), x_{k+1} - x_k\right\rangle + \left(\frac{L}{2} - \frac{1}{2\eta}\right)\|x_{k+1} - x_k\|^2\right)\Bigg] \\ &\leq \mathbb{E}_k\Bigg[- \left(\frac{L}{2}\|x_{k+1} - x_k\|^2 + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + g(x_{k+1})\right) \\ &\qquad + \left(\frac{1}{2L\lambda}\|\nabla f(x_k) - \widetilde{\nabla} f(x_k)\|^2 + \left(\frac{L(\lambda+1)}{2} - \frac{1}{2\eta}\right)\|x_{k+1} - x_k\|^2\right)\Bigg] \\ &\leq \mathbb{E}_k\left[f(x_k) - F(x_{k+1}) + \left(\frac{1}{2L\lambda}\|\nabla f(x_k) - \widetilde{\nabla} f(x_k)\|^2 + \left(\frac{L(\lambda+1)}{2} - \frac{1}{2\eta}\right)\|x_{k+1} - x_k\|^2\right)\right]\end{aligned}$$

The first inequality is Young's: $\langle a, b\rangle \leq \frac{1}{2\epsilon}\|a\|^2 + \frac{\epsilon}{2}\|b\|^2$, where we set $\epsilon = \lambda L$ for some positive constant λ whose value we determine later. The second inequality follows from the fact that f has an L-Lipschitz continuous gradient. □

We are now prepared to prove our general inequality.

Lemma 12 (Restatement of Lemma 1). *Suppose g is μ-strongly convex with $\mu \geq 0$, and set the bias parameter $\theta \geq 1$. Let $\lambda > 0$ be a constant whose value we determine later and the operator $\widetilde{\nabla} \equiv \widetilde{\nabla}_{\text{B-SAGA}}$ or $\widetilde{\nabla}_{\text{B-SVRG}}$. The following inequality holds:*

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

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

$$\begin{aligned}& \frac{\eta}{\theta}(f(x_k) - f(x^*)) + \frac{\eta}{n}\left(1 - \frac{1}{\theta}\right)\left(\sum_{i=1}^n \nabla f_i(\varphi_k^i) - \nabla f(x^*)\right) \\ \leq & \frac{\eta}{\theta}\langle \nabla f(x_k), x_k - x^*\rangle + \frac{\eta}{n}\left(1 - \frac{1}{\theta}\right)\sum_{i=1}^n \langle \nabla f_i(\varphi_k^i), \varphi_k^i - x^*\rangle \\ = & \frac{\eta}{\theta}\langle \nabla f(x_k), x_k - x^*\rangle + \frac{\eta}{n}\left(1 - \frac{1}{\theta}\right)\sum_{i=1}^n \langle \nabla f_i(\varphi_k^i), x_k - x^*\rangle \\ & + \frac{\eta}{n}\left(1 - \frac{1}{\theta}\right)\sum_{i=1}^n \langle \nabla f_i(\varphi_k^i), \varphi_k^i - x_k\rangle.\end{aligned}$$

We now bound the first line on the right using Lemma 10. The expected value of our gradient estimate is

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

with the understanding that in the case $\widetilde{\nabla} \equiv \widetilde{\nabla}_{\text{SVRG}}$, the points $\varphi_k^i = \varphi_s$ for all i when k is in epoch s. Therefore,

$$\begin{aligned}
&\frac{\eta}{\theta}\langle \nabla f(x_k), x_k - x^*\rangle + \frac{\eta}{n}\left(1 - \frac{1}{\theta}\right)\sum_{i=1}^{n}\langle \nabla f_i(\varphi_k^i), x_k - x^*\rangle \\
=&\mathbb{E}_k\left[\eta\langle \widetilde{\nabla} f(x_k), x_k - x^*\rangle\right] \\
=&\mathbb{E}_k\left[\eta\langle \widetilde{\nabla} f(x_k), x_k - x_{k+1}\rangle + \eta\langle \widetilde{\nabla} f(x_k), x_{k+1} - x^*\rangle\right] \\
\leq&\mathbb{E}_k\Big[\eta\langle \widetilde{\nabla} f(x_k), x_k - x_{k+1}\rangle - \frac{1}{2}\|x_k - x_{k+1}\|^2 + \frac{1}{2}\|x_k - x^*\|^2 - \frac{1}{2}\|x_{k+1} - x^*\|^2 \\
&\quad - \eta g(x_{k+1}) + \eta g(x^*)\Big],
\end{aligned}$$

The inequality is due to Lemma 10 with $u = x^*$. Combining these two inequalities, we have shown

$$\begin{aligned}
&\frac{\eta}{\theta}(f(x_k) - f(x^*)) + \frac{\eta}{n}\left(1 - \frac{1}{\theta}\right)\left(\sum_{i=1}^{n}\nabla f_i(\varphi_k^i) - \nabla f(x^*)\right) \\
\leq&\mathbb{E}_k\Big[\eta\langle \widetilde{\nabla} f(x_k), x_k - x_{k+1}\rangle - \frac{1}{2}\|x_k - x_{k+1}\|^2 - \eta g(x_{k+1}) + \eta g(x^*) \qquad (7)\\
&\quad + \frac{1}{2}\|x_k - x^*\|^2 - \frac{1}{2}\|x_{k+1} - x^*\|^2 + \frac{\eta}{n}\left(1 - \frac{1}{\theta}\right)\sum_{i=1}^{n}\langle \nabla f_i(\varphi_k^i), \varphi_k^i - x_k\rangle\Big].
\end{aligned}$$

To complete the proof, we use Lemma 11 to bound the terms on the top line.

$$\begin{aligned}
&= \mathbb{E}_k\left[\eta\langle \widetilde{\nabla} f(x_k), x_k - x_{k+1}\rangle - \frac{1}{2}\|x_k - x_{k+1}\|^2 - \eta g(x_{k+1})\right] \\
&= -\eta\mathbb{E}_k\left[\langle \widetilde{\nabla} f(x_k), x_{k+1} - x_k\rangle + \frac{1}{2}\|x_k - x_{k+1}\|^2 + \eta g(x_{k+1})\right] \\
&= \eta\mathbb{E}_k\left[\text{Prog}(x_{k+1})\right] \\
&\leq \eta f(x_k) - \eta\mathbb{E}_k\left[F(x_{k+1})\right] + \frac{\eta}{2L\lambda}\mathbb{E}_k\|\widetilde{\nabla} f(x_k) - \nabla f(x_k)\|^2 + \left(\frac{\eta L(\lambda+1)}{2} - \frac{1}{2}\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2
\end{aligned}$$

Combining this bound with (D) and rearranging terms, we have shown that

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

Finally, we use Lemma 7 to bound the final term, yielding the desired inequality

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

□

E Proofs for B-SAGA

All we require before proving our main result is a bound on the MSE of our gradient estimator and a way to ensure these terms telescope. The next lemma is the former.

Lemma 13 (**Restatement of Lemma 2**). *The MSE of the SAGA gradient estimator satisfies*

$$\begin{aligned}\mathbb{E}_k\|\widetilde{\nabla}_{\text{SAGA}}f(x_k)-\nabla f(x_k)\|^2 \leq \frac{L^2}{n\theta^2}\sum_{i=1}^n\|x_k-\varphi_k^i\|^2 \\ +\left(1-\frac{2}{\theta}\right)\left\|\nabla f(x_k)-\frac{1}{n}\sum_{i=1}^n\nabla f_i(\varphi_k^i)\right\|^2\end{aligned}$$

Proof. The proof amounts to computing the expectation and applying the Lipschitz continuity of ∇f_i.

$$\begin{aligned}&\mathbb{E}_k\|\widetilde{\nabla}_{\text{SAGA}}f(x_k)-\nabla f(x_k)\|^2\\ =&\mathbb{E}_k\left\|\left(\frac{1}{\theta}\right)\left(\nabla f_{j_k}(x_k)-\nabla f_{j_k}(\varphi_k^{j_k})\right)-\nabla f(x_k)-\frac{1}{n}\sum_{i=1}^n\nabla f_i(\varphi_k^i)\right\|^2\\ =&\frac{1}{\theta^2}\mathbb{E}_k\left\|\nabla f_{j_k}(x_k)-\nabla f_{j_k}(\varphi_k^{j_k})\right\|^2+\left\|\nabla f(x_k)-\frac{1}{n}\sum_{i=1}^n\nabla f_i(\varphi_k^i)\right\|^2\\ &-\frac{2}{\theta}\mathbb{E}_k\left\langle\nabla f_{j_k}(x_k)-\nabla f_{j_k}(\varphi_k^{j_k}),\nabla f(x_k)-\frac{1}{n}\sum_{i=1}^n\nabla f_i(\varphi_k^i)\right\rangle\\ =&\frac{1}{n\theta^2}\sum_{i=1}^n\left\|\nabla f_i(x_k)-\nabla f_i(\varphi_k^i)\right\|^2+\left(1-\frac{2}{\theta}\right)\left\|\nabla f(x_k)-\frac{1}{n}\sum_{i=1}^n\nabla f_i(\varphi_k^i)\right\|^2\\ \leq&\frac{L^2}{n\theta^2}\sum_{i=1}^n\left\|x_k-\varphi_k^i\right\|^2+\left(1-\frac{2}{\theta}\right)\left\|\nabla f(x_k)-\frac{1}{n}\sum_{i=1}^n\nabla f_i(\varphi_k^i)\right\|^2.\end{aligned}$$

□

We see that for θ close to one, this bound is of order $\mathcal{O}(\frac{1}{\theta^2})$, so increasing θ decreases our bound on the MSE. However, as θ becomes large, the gradient estimate does not approximate $\nabla f(x_k)$ as well, and the bound in Lemma 13 reflects this. The next lemma allows us to prove that the terms in the bound of Lemma 13 telescope.

Lemma 14. *Let $c \geq 1$ and $\delta > 0$ be constants whose value we determine later. The following inequality holds:*

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\left\|x_k-\varphi_k^i\right\|^2\\ \leq&-c(1-\delta)\sum_{i=1}^n\mathbb{E}_k\left\|x_{k+1}-\varphi_{k+1}^i\right\|^2+c\left(1-\frac{c-1}{cn}\right)\sum_{i=1}^n\left\|x_k-\varphi_k^i\right\|^2\\ &+c(\delta^{-1}-1)L^2n\mathbb{E}_k\left\|x_{k+1}-x_k\right\|^2\end{aligned}$$

Proof. Computing expectations, we see that

$$\begin{aligned}\sum_{i=1}^n\mathbb{E}_k\left\|x_k-\varphi_{k+1}^i\right\|^2&=\|x_k-\varphi_{k+1}^{j_k}\|^2+\mathbb{E}_k\left[\sum_{i\neq j_k}\left\|x_k-\varphi_k^i\right\|^2\right]\\ &=\left(1-\frac{1}{n}\right)\sum_{i=1}^n\left\|x_k-\varphi_k^i\right\|^2,\end{aligned}$$

where we have applied the update rule $\varphi_{k+1}^{j_k} = x_k$. Multiplying this equality by a constant $c > 0$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left\|x_k - \varphi_k^i\right\|^2 \\
={}& - c\sum_{i=1}^{n}\mathbb{E}_k\left\|x_k - \varphi_{k+1}^i\right\|^2 + c\left(1 - \frac{c-1}{cn}\right)\sum_{i=1}^{n}\left\|x_k - \varphi_k^i\right\|^2 .
\end{aligned}$$

Using the inequality $-\|u - w\|^2 \leq -(1-\delta)\|u - v\|^2 + (\delta^{-1} - 1)\|v - w\|^2$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left\|x_k - \varphi_k^i\right\|^2 \\
={}& - c\sum_{i=1}^{n}\mathbb{E}_k\left\|x_k - \varphi_{k+1}^i\right\|^2 + c\left(1 - \frac{c-1}{cn}\right)\sum_{i=1}^{n}\left\|x_k - \varphi_k^i\right\|^2 \\
\leq{}& - c(1-\delta)\sum_{i=1}^{n}\mathbb{E}_k\left\|x_{k+1} - \varphi_{k+1}^i\right\|^2 + c\left(1 - \frac{c-1}{cn}\right)\sum_{i=1}^{n}\left\|x_k - \varphi_k^i\right\|^2 \\
&+ c(\delta^{-1} - 1)\sum_{i=1}^{n}\mathbb{E}_k\left\|x_{k+1} - x_k\right\|^2 .
\end{aligned}$$

□

We are now prepared to prove our main result. For now, consider the non-strongly convex case with $\mu = 0$.

E.1 Convex

Theorem 15 (Restatement of Theorem 3, Part 1). *In Algorithm 1, set* $\eta = \frac{\theta}{4Ln^2(\theta-1)+L(\theta+3\sqrt{2}n)}$ *for* $\theta \in [1, 2)$*, and set* $\eta = \frac{\theta}{4Ln^2(\theta-1)+3\sqrt{2}Ln(\theta-1)+L\theta}$ *for* $\theta \geq 2$*. After* T *iterations of Algorithm 1, we have the following bound on the suboptimality of the average iterate* $\overline{x} \overset{def}{=} \frac{1}{T}\sum_{k=1}^{T} x_k$ *:*

$$\mathbb{E}\left[F(\overline{x}) - F(x^*)\right] \leq \begin{cases} \frac{2Ln^2(\theta-1)+\frac{L}{2}(\theta+3\sqrt{2}n)}{\theta T}\|x_0 - x^*\|^2 & \theta \in [1,2), \\ \frac{2Ln^2(\theta-1)+\frac{3\sqrt{2}}{2}Ln(\theta-1)+L\theta}{\theta T}\|x_0 - x^*\|^2 & \theta \geq 2. \end{cases}$$

Proof. From Lemma 12, we have

$$\begin{aligned}
0 \leq{}& - \eta\mathbb{E}_k\left[F(x_{k+1}) - F(x^*)\right] + \frac{\eta}{2L\lambda}\mathbb{E}_k\|\widetilde{\nabla}_{\text{SAGA}}f(x_k) - \nabla f(x_k)\|^2 - \frac{1}{2}\mathbb{E}_k\|x_{k+1} - x^*\|^2 \\
&+ \frac{1}{2}\|x_k - x^*\|^2 + \eta\left(1 - \frac{1}{\theta}\right)\sum_{i=1}^{n}\|x_k - \varphi_k^i\|^2 \\
&+ \left(\frac{\eta L(\lambda+1)}{2} - \frac{1}{2}\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2.
\end{aligned}$$

Using our bound on the MSE, we have

$$\begin{aligned}
0 \leq{}& - \eta\mathbb{E}_k\left[F(x_{k+1}) - F(x^*)\right] + \frac{\eta}{2L\lambda}\left(1 - \frac{2}{\theta}\right)\left\|\nabla f(x_k) - \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\varphi_k^i)\right\|^2 \\
&- \frac{1}{2}\mathbb{E}_k\|x_{k+1} - x^*\|^2 + \frac{1}{2}\|x_k - x^*\|^2 + \left(\frac{\eta L(\lambda+1)}{2} - \frac{1}{2}\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2 \\
&+ \frac{\eta L}{2n}\left(1 - \frac{1}{\theta} + \frac{1}{\lambda\theta^2}\right)\sum_{i=1}^{n}\|x_k - \varphi_k^i\|^2.
\end{aligned} \tag{8}$$

We now consider the following two cases: $\theta \in [1, 2)$ and $\theta \geq 2$.

Case 1. Suppose $\theta \in [1,2)$. In this case, $1-\frac{2}{\theta} \le 0$, so we can drop the second term in (E.1). Applying Lemma 14 produces the inequality

$$\begin{aligned}0 \le &- \eta\mathbb{E}_k\left[F(x_{k+1})-F(x^*)\right]\\&-\frac{1}{2}\mathbb{E}_k\|x_{k+1}-x^*\|^2+\frac{1}{2}\|x_k-x^*\|^2+\left(\frac{\eta L(\lambda+1)}{2}-\frac{1}{2}\right)\|x_{k+1}-x_k\|^2\\&+\frac{c\eta L}{2}\left(1-\frac{1}{\theta}+\frac{1}{\lambda\theta^2}\right)\left(1-\frac{c-1}{cn}\right)\sum_{i=1}^n\|x_k-\varphi_k^i\|^2\\&-\frac{c\eta L}{2}\left(1-\frac{1}{\theta}+\frac{1}{\lambda\theta^2}\right)(1-\delta)\sum_{i=1}^n\mathbb{E}_k\|x_{k+1}-\varphi_{k+1}^i\|^2\\&+\frac{c\eta Ln\delta^{-1}}{2}\left(1-\frac{1}{\theta}+\frac{1}{\lambda\theta^2}\right)\mathbb{E}_k\|x_{k+1}-x_k\|^2.\end{aligned}$$

Define the Lyapunov functional

$$E_k^1 \overset{\text{def}}{=} \frac{1}{2}\|x_k-x^*\|^2+\frac{c\eta L}{2}\left(1-\frac{1}{\theta}+\frac{1}{\lambda\theta^2}\right)\left(1-\frac{c-1}{cn}\right)\sum_{i=1}^n\|x_k-\varphi_k^i\|^2. \tag{9}$$

Setting $\delta = \frac{c-1}{cn}$, we have shown that

$$\begin{aligned}\mathbb{E}_k\left[E_{k+1}^1\right]-E_k^1 \le &-\eta\mathbb{E}_k[F(x_{k+1})-F(x^*)]\\&+\left(\frac{\eta L(\lambda+1)}{2}-\frac{1}{2}+\frac{c^2\eta Ln^2}{2(c-1)}\left(1-\frac{1}{\theta}+\frac{1}{\lambda\theta^2}\right)\right)\mathbb{E}_k\|x_{k+1}-x_k\|^2\end{aligned}$$

We choose $c=2$ and $\lambda = \frac{2\sqrt{2}n}{\theta}$ to minimize the coefficient of the final term. This term is non-positive as long as

$$\eta \le \frac{\theta}{4Ln^2(\theta-1)+L(\theta+3\sqrt{2}n)}.$$

With these parameter choices, $\mathbb{E}_k\left[E_{k+1}^1\right] \le E_k^1 - \eta\mathbb{E}_k[F(x_{k+1})-F(x^*)]$. Telescoping this inequality from $k=0$ to $k=T-1$ and chaining the conditional expectations, we have

$$\begin{aligned}\mathbb{E}\left[\eta\sum_{k=1}^T(F(x_k)-F(x^*))\right] &\le -\mathbb{E}[E_T]+E_0\\&\le \frac{1}{2}\|x_0-x^*\|^2+\left(1-\frac{1}{\theta}+\frac{1}{\lambda\theta^2}\right)\left(1-\frac{1}{2n}\right)\sum_{i=1}^n\|x_k-\varphi_k^i\|^2.\end{aligned}$$

Because $x_0=\varphi_0^i$ for all i, the final term on the right is equal to zero. Define $\overline{x} \overset{\text{def}}{=} \frac{1}{T}\sum_{k=1}^T x_k$. The convexity of F combined with the inequality above implies

$$\eta T\mathbb{E}\left[F(\overline{x})-F(x^*)\right] \le \frac{1}{2}\|x_0-x^*\|^2.$$

Choosing η maximally, $\eta = \frac{\theta}{4Ln^2(\theta-1)+L(\theta+3\sqrt{2}n)}$, completes the proof.

Case 2. Now suppose that $\theta \ge 2$. Recall the bound from (E.1):

$$\begin{aligned}0 \le &-\eta\mathbb{E}_k\left[F(x_{k+1})-F(x^*)\right]+\frac{\eta}{2L\lambda}\left(1-\frac{2}{\theta}\right)\left\|\nabla f(x_k)-\frac{1}{n}\sum_{i=1}^n\nabla f_i(\varphi_k^i)\right\|^2\\&-\frac{1}{2}\mathbb{E}_k\|x_{k+1}-x^*\|^2+\frac{1}{2}\|x_k-x^*\|^2+\left(\frac{\eta L(\lambda+1)}{2}-\frac{1}{2}\right)\mathbb{E}_k\|x_{k+1}-x_k\|^2\\&+\frac{\eta L}{2n}\left(1-\frac{1}{\theta}+\frac{1}{\lambda\theta^2}\right)\sum_{i=1}^n\|x_k-\varphi_k^i\|^2.\end{aligned}$$

Because $\theta \geq 2$, Jensen's inequality and the Lipschitz continuity of ∇f_i gives us

$$\begin{aligned}
0 \leq& \left(1 - \frac{2}{\theta}\right) \left\| \nabla f(x_k) - \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\varphi_k^i) \right\|^2 \\
\leq& \left(1 - \frac{2}{\theta}\right) \frac{1}{n}\sum_{i=1}^{n} \left\| \nabla f_i(x_k) + \nabla f_i(\varphi_k^i) \right\|^2 \\
\leq& \frac{L^2}{n}\left(1 - \frac{2}{\theta}\right) \sum_{i=1}^{n} \|x_k - \varphi_k^i\|^2.
\end{aligned}$$

We now have the inequality

$$\begin{aligned}
0 \leq& -\eta \mathbb{E}_k\left[F(x_{k+1}) - F(x^*)\right] - \frac{1}{2}\mathbb{E}_k\|x_{k+1} - x^*\|^2 + \frac{1}{2}\|x_k - x^*\|^2 \\
&+ \left(\frac{\eta L(\lambda+1)}{2} - \frac{1}{2}\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2 + \frac{\eta L}{2}\left(1 + \frac{1}{\lambda} - \frac{1+\frac{2}{\lambda}}{\theta} + \frac{1}{\lambda\theta^2}\right)\sum_{i=1}^{n}\|x_k - \varphi_k^i\|^2,
\end{aligned} \tag{10}$$

and we can proceed just as in Case 1. Applying Lemma 14,

$$\begin{aligned}
0 \leq& -\eta \mathbb{E}_k\left[F(x_{k+1}) - F(x^*)\right] - \frac{1}{2}\mathbb{E}_k\|x_{k+1} - x^*\|^2 + \frac{1}{2}\|x_k - x^*\|^2 \\
&+ \left(\frac{\eta L(\lambda+1)}{2} - \frac{1}{2}\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2 \\
&+ \frac{c\eta L}{2}\left(1 + \frac{1}{\lambda} - \frac{1+\frac{2}{\lambda}}{\theta} + \frac{1}{\lambda\theta^2}\right)\left(1 - \frac{c-1}{cn}\right)\sum_{i=1}^{n}\|x_k - \varphi_k^i\|^2 \\
&- \frac{c\eta L}{2}\left(1 + \frac{1}{\lambda} - \frac{1+\frac{2}{\lambda}}{\theta} + \frac{1}{\lambda\theta^2}\right)(1-\delta)\sum_{i=1}^{n}\mathbb{E}_k\|x_{k+1} - \varphi_{k+1}^i\|^2 \\
&+ \frac{c\eta L n\delta^{-1}}{2}\left(1 + \frac{1}{\lambda} - \frac{1+\frac{2}{\lambda}}{\theta} + \frac{1}{\lambda\theta^2}\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2.
\end{aligned}$$

Define the Lyapunov functional

$$E_k^2 \overset{\text{def}}{=} \frac{1}{2}\|x_k - x^*\|^2 + \frac{c\eta L}{2}\left(1 + \frac{1}{\lambda} - \frac{1+\frac{2}{\lambda}}{\theta} + \frac{1}{\lambda\theta^2}\right)\left(1 - \frac{c-1}{cn}\right)\sum_{i=1}^{n}\|x_k - \varphi_k^i\|^2. \tag{11}$$

Setting $\delta = \frac{c-1}{cn}$, we have

$$\begin{aligned}
\mathbb{E}_k\left[E_{k+1}^2\right] - E_k^2 \leq& -\eta\mathbb{E}_k[F(x_{k+1}) - F(x^*)] \\
&+ \left(\frac{\eta L(\lambda+1)}{2} - \frac{1}{2} + 2n^2 L\eta\left(1 + \frac{1}{\lambda} - \frac{1+\frac{2}{\lambda}}{\theta} + \frac{1}{\lambda\theta^2}\right)\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2.
\end{aligned}$$

We set $c = 2$ as in Case 1, but we choose $\lambda = \frac{2\sqrt{2}n(\theta-1)}{\theta}$ differently. These choices minimise the coefficient of the final term above. With these parameter choices, the terms in the round brackets are non-positive as long as

$$\eta \leq \frac{\theta}{4Ln^2(\theta-1) + 3\sqrt{2}Ln(\theta-1) + L\theta}.$$

Choosing η maximally and following the same telescoping procedure as in Case 1, we are done.

□

E.2 Strongly convex

Theorem 16 (Restatement of Theorem 3, Part 2). *In Algorithm 1, set $\eta = \frac{\theta}{4Ln^2(\theta-1)+L(\theta+3\sqrt{2}n)}$ for $\theta \in [1,2)$, and set $\eta = \frac{\theta}{4Ln^2(\theta-1)+3\sqrt{2}Ln(\theta-1)+L\theta}$ for $\theta \geq 2$. If g is μ-strongly convex, after T*

iterations, Algorithm 1 produces an iterate satisfying

$$\mathbb{E}\|x_T - x^*\|^2 \leq \begin{cases} \left(1 + \frac{\mu\theta}{8Ln^2(\theta-1)+2L(\theta+3\sqrt{2}n)}\right)^{-T} \|x_0 - x^*\|^2 & \text{for } \theta \in [1,2) \\ \left(1 + \frac{\mu\theta}{8Ln^2(\theta-1)+6\sqrt{2}Ln(\theta-1)+2L\theta}\right)^{-T} \|x_0 - x^*\|^2 & \text{for } \theta \geq 2. \end{cases}$$

Proof. Recall inequality (E.1) from our analysis of the non-strongly convex case.

$$\begin{aligned} 0 \leq & -\eta \mathbb{E}_k \left[F(x_{k+1}) - F(x^*)\right] + \frac{\eta}{2L\lambda}\left(1 - \frac{2}{\theta}\right)\left\|\nabla f(x_k) - \frac{1}{n}\sum_{i=1}^n \nabla f_i(\varphi_k^i)\right\|^2 \\ & - \frac{1}{2}\mathbb{E}_k\|x_{k+1} - x^*\|^2 + \frac{1}{2}\|x_k - x^*\|^2 + \left(\frac{\eta L(\lambda+1)}{2} - \frac{1}{2}\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2 \\ & + \frac{\eta L}{2}\left(1 - \frac{1}{\theta} + \frac{1}{\lambda\theta^2}\right)\sum_{i=1}^n \|x_k - \varphi_k^i\|^2. \end{aligned}$$

We drop the non-positive term $-\eta\left[F(x_{k+1}) - F(x^*)\right]$. As before, we divide our analysis into two cases.

Case 1. Suppose $\theta \in [1,2)$. In this case, $1 - \frac{2}{\theta} \leq 0$, so we can drop the first term in (E.1). Applying Lemma 14 produces the inequality

$$\begin{aligned} 0 \leq & -\frac{1+\mu\eta}{2}\mathbb{E}_k\|x_{k+1} - x^*\|^2 + \frac{1}{2}\|x_k - x^*\|^2 + \left(\frac{\eta(L+\lambda)}{2} - \frac{1}{2}\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2 \\ & - \frac{c\eta L}{2}\left(1 - \frac{1}{\theta} + \frac{1}{\lambda\theta^2}\right)(1-\delta)\sum_{i=1}^n \mathbb{E}_k\|x_{k+1} - \varphi_{k+1}^i\|^2 \\ & + \frac{c\eta L}{2}\left(1 - \frac{1}{\theta} + \frac{1}{\lambda\theta^2}\right)\left(1 - \frac{c-1}{cn}\right)\sum_{i=1}^n \|x_k - \varphi_k^i\|^2 \\ & + \frac{cn\delta^{-1}L\eta}{2}\left(1 - \frac{1}{\theta} + \frac{1}{\lambda\theta^2}\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2. \end{aligned}$$

With the Lyapunov functional E_k^1 defined in (E.1), we have shown

$$\begin{aligned} & (1+\kappa_1)\mathbb{E}_k\left[E_{k+1}^1\right] - E_k^1 \\ \leq & \left(\kappa_1 - \frac{\mu\eta}{2}\right)\mathbb{E}_k\|x_{k+1} - x^*\|^2 \\ & + c\eta\left(1 - \frac{1}{\theta} + \frac{1}{\lambda\theta^2}\right)\left((1+\kappa_1)\left(1 - \frac{(c-1)}{cn}\right) - (1-\delta)\right)\sum_{i=1}^n \mathbb{E}_k\|x_{k+1} - \varphi_{k+1}^i\|^2 \\ & + \left(\frac{\eta L(\lambda+1)}{2} - \frac{1}{2} + \frac{cn\delta^{-1}L\eta}{2}\left(1 - \frac{1}{\theta} + \frac{1}{\lambda\theta^2}\right)\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2. \end{aligned}$$

We would like to choose parameters so that κ_1 is maximised. For ease of exposition, we only approximately maximise κ_1, and set the parameters $\kappa_1 = \frac{\mu\eta}{2}$ and $\eta = \frac{\theta}{4Ln^2(\theta-1)+L(\theta+3\sqrt{2}n)}$. This ensures that $\kappa_1 \leq \frac{1}{8n}$, so we can choose $c = 2$, $\lambda = \frac{\sqrt{6}n}{\theta}$, and $\delta = \frac{1}{3n}$ to make the remaining terms non-positive. Chaining this inequality and the expectations over the iterations $k = 0$ to $k = T-1$, we have

$$\mathbb{E}\left[E_T^1\right] \leq \left(1 + \frac{\mu\theta}{8Ln^2(\theta-1) + 2L(\theta+3\sqrt{2}n)}\right)^{-T} E_0^1,$$

which implies

$$\mathbb{E}\|x_T - x^*\|^2 \leq \left(1 + \frac{\mu\theta}{8Ln^2(\theta-1) + 2L(\theta+3\sqrt{2}n)}\right)^{-T} \|x_0 - x^*\|^2.$$

Case 2. For $\theta \geq 2$, we apply the bound

$$\frac{\eta}{2L\lambda}\left(1-\frac{2}{\theta}\right)\left\|\nabla f(x_k)-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\varphi_k^i)\right\|^2 \leq \frac{\eta L}{2\lambda n}\left(1-\frac{2}{\theta}\right)\sum_{i=1}^{n}\left\|x_k-\varphi_k^i\right\|^2$$

to the inequality in (E.1). Following the same procedure as in Case 1 and using the Lyapunov functional E_k^2 defined in (E.1), we have that

$$\begin{aligned}
&(1+\kappa_2)\,\mathbb{E}_k\left[E_{k+1}^1\right]-E_k^1\\
&\leq \left(\kappa_2-\frac{\mu\eta}{2}\right)\mathbb{E}_k\|x_{k+1}-x^*\|^2\\
&\quad + c\eta\left(1+\frac{1}{\lambda}-\frac{1+\frac{2}{\lambda}}{\theta}+\frac{1}{\lambda\theta^2}\right)\left((1+\kappa_2)\left(1-\frac{(c-1)}{cn}\right)-(1-\delta)\right)\sum_{i=1}^{n}\mathbb{E}_k\|x_{k+1}-\varphi_{k+1}^i\|^2\\
&\quad + \left(\frac{\eta L(\lambda+1)}{2}-\frac{1}{2}+\frac{cn\delta^{-1}L\eta}{2}\left(1+\frac{1}{\lambda}-\frac{1+\frac{2}{\lambda}}{\theta}+\frac{1}{\lambda\theta^2}\right)\right)\mathbb{E}_k\|x_{k+1}-x_k\|^2.
\end{aligned}$$

We set $\kappa_2 = \frac{\mu\eta}{2}$ and $\eta = \frac{\theta}{4Ln^2(\theta-1)+3\sqrt{2}Ln(\theta-1)+L\theta}$. Notice that this implies $\kappa_2 \leq \frac{1}{8n}$. We choose $\delta = \frac{1}{3n}$, $\lambda = \frac{\sqrt{6}n(\theta-1)}{\theta}$, and $c = 2$ to make the remaining term non-positive. Chaining the resulting inequality over the iterations $k = 0$ to $k = T-1$, we obtain

$$\mathbb{E}\left[E_T^2\right] \leq \left(1+\frac{\mu\theta}{8Ln^2(\theta-1)+6\sqrt{2}Ln(\theta-1)+2L\theta}\right)^{-T}E_0^2,$$

which implies

$$\mathbb{E}\|x_T-x^*\|^2 \leq \left(1+\frac{\mu\theta}{8Ln^2(\theta-1)+6\sqrt{2}Ln(\theta-1)+2L\theta}\right)^{-T}\|x_0-x^*\|^2.$$

□

E.3 Non-convex

Theorem 17 (Restatement of Theorem 5). *In Algorithm 1, set $\eta = \frac{\theta}{2Ln}$ for $\theta \leq 2$, and set $\eta = \frac{\theta}{2Ln(\theta-1)}$ for $\theta \geq 2$. After T steps, Algorithm 1 achieves the following bound on the norm of the generalised gradient:*

$$\mathbb{E}\|\mathcal{G}_\eta(x_\alpha)\|^2 \leq \begin{cases} \frac{4Ln(F(x_0)-F(x^*))}{\theta\left(1-\frac{\theta}{n}\right)T} & \text{for } 0<\theta<2,\\ \frac{4Ln(\theta-1)(F(x_0)-F(x^*))}{\theta\left(1-\frac{\theta}{n(\theta-1)}\right)T} & \text{for } \theta \geq 2.\end{cases}$$

Proof. Define

$$\hat{x}_{k+1} \stackrel{\text{def}}{=} \text{prox}_{\eta g}\left(x_k-\nabla f(x_k)\right). \tag{12}$$

We can interpret $\hat{x}_{k+1}$ as the iterate that would be produced from x_k if the full gradient $\nabla f(x_k)$ were available. Because it does not rely on a stochastic gradient at x_k, it is independent of j_k. The first two steps of our proof follow the proof of Theorem 5 in [20]. Applying Lemma 9 with $y = \hat{x}$, $z = x_k$, and $d = \nabla f(x_k)$, we have

$$\mathbb{E}_k\left[F(\hat{x}_{k+1})\right] \leq \mathbb{E}_k\left[F(x_k)+\left(\frac{L}{2}-\frac{1}{2\eta}\right)\|\hat{x}_{k+1}-x_k\|^2-\frac{1}{2\eta}\|\hat{x}_{k+1}-x_k\|^2\right].$$

Furthermore, applying Lemma 2 with $y = \hat{x}_{k+1}$, $z = x_k$, and $d = \nabla f(x_k)$, we also have the inequality

$$\mathbb{E}_k\left[F(x_{k+1})\right] \leq \mathbb{E}_k\Bigg[F(\hat{x}_{k+1})+\langle x_{k+1}-\hat{x}_{k+1},\nabla f(x_k)-\widetilde{\nabla}f(x_k)\rangle+\left(\frac{L}{2}-\frac{1}{2\eta}\right)\|x_{k+1}-x_k\|^2$$

$$+ \left(\frac{L}{2} + \frac{1}{2\eta}\right) \|\hat{x}_{k+1} - x_k\|^2 - \frac{1}{2\eta}\|x_{k+1} - \hat{x}_{k+1}\|^2\Bigg].$$

Adding these two inequalities together gives

$$\begin{aligned}\mathbb{E}_k\left[F(x_{k+1})\right] &\leq \mathbb{E}_k\Bigg[F(x_k) + \left(L - \frac{1}{2\eta}\right)\|\hat{x}_{k+1} - x_k\|^2 + \left(\frac{L}{2} - \frac{1}{2\eta}\right)\|x_{k+1} - x_k\|^2 \\ &\qquad - \frac{1}{2\eta}\|x_{k+1} - \hat{x}_{k+1}\|^2 + \langle x_{k+1} - \hat{x}_{k+1}, \nabla f(x_k) - \widetilde{\nabla} f(x_k)\rangle\Bigg] \\ &\leq \mathbb{E}_k\Bigg[F(x_k) + \left(L - \frac{1}{2\eta}\right)\|\hat{x}_{k+1} - x_k\|^2 + \left(\frac{L}{2} - \frac{1}{2\eta}\right)\|x_{k+1} - x_k\|^2 \\ &\qquad + \frac{\eta}{2}\|\nabla f(x_k) - \widetilde{\nabla} f(x_k)\|^2\Bigg].\end{aligned}$$

The inequality on the last line is an application of Young's inequality. We bound the final term using Lemma 13. Adding these two inequalities together gives

$$\begin{aligned}0 &\leq -\mathbb{E}_k\left[F(x_{k+1})\right] + F(x_k) + \left(L - \frac{1}{2\eta}\right)\|\hat{x}_{k+1} - x_k\|^2 + \left(\frac{L}{2} - \frac{1}{2\eta}\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2 \\ &\quad + \frac{\eta L^2}{2n\theta^2}\sum_{i=1}^n \|x_k - \varphi_k^i\|^2 + \frac{\eta}{2}\left(1 - \frac{2}{\theta}\right)\left\|\nabla f(x_k) - \frac{1}{n}\sum_{i=1}^n \nabla f_i(\varphi_k^i)\right\|^2. \qquad (13)\end{aligned}$$

We now split our analysis into two cases.

Case 1. Suppose $\theta \in (0, 2)$. The term

$$\frac{1}{2Ln}\left(1 - \frac{2}{\theta}\right)\left\|\nabla f(x_k) - \frac{1}{n}\sum_{i=1}^n \nabla f_i(\varphi_k^i)\right\|^2$$

is then non-positive, so we can drop it from the inequality in (E.3). Applying Lemma 14 to the remaining term from the MSE bound gives

$$\begin{aligned}0 &\leq -\mathbb{E}_k\left[F(x_{k+1})\right] - F(x_k) + \left(L - \frac{1}{2\eta}\right)\|\hat{x}_{k+1} - x_k\|^2 \\ &\quad - \frac{c\eta L^2}{2\theta^2}(1 - \delta)\sum_{i=1}^n \mathbb{E}_k\left\|x_{k+1} - \varphi_{k+1}^i\right\|^2 + \frac{c\eta L^2}{2\theta^2}\left(1 - \frac{c-1}{cn}\right)\sum_{i=1}^n \left\|x_k - \varphi_k^i\right\|^2 \\ &\quad + \left(\frac{cn\delta^{-1}\eta L^2}{2\theta^2} + \frac{L}{2} - \frac{1}{2\eta}\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2.\end{aligned}$$

We choose $\delta = \frac{c-1}{cn}$ so that the terms in the second line telescope over several iterations. We must also choose c and η so that the final term is non-positive. The coefficient of this term is

$$\frac{c^2n^2\eta L^2}{2\theta^2(c-1)} + \frac{L}{2} - \frac{1}{2\eta},$$

so the minimising c is $c = 2$, and the resulting expression is non-positive as long as

$$\eta \leq \frac{\sqrt{16n^2\theta^2 + \theta^4}}{8Ln^2} - \frac{\theta^2}{8Ln^2}. \qquad (14)$$

With these parameter choices, we are left with the inequality

$$\begin{aligned}0 &\leq -\mathbb{E}_k\left[F(x_{k+1})\right] - F(x_k) + \left(L - \frac{1}{2\eta}\right)\|\hat{x}_{k+1} - x_k\|^2 \\ &\quad - \frac{\eta L^2}{\theta^2}\left(1 - \frac{1}{2n}\right)\sum_{i=1}^n \mathbb{E}_k\left\|x_{k+1} - \varphi_{k+1}^i\right\|^2 + \frac{\eta L^2}{\theta^2}\left(1 - \frac{1}{2n}\right)\sum_{i=1}^n \left\|x_k - \varphi_k^i\right\|^2.\end{aligned}$$

Summing this inequality from $k = 0$ to $k = T - 1$ and chaining the conditional expectations, we have

$$\begin{aligned}
0 \leq -\mathbb{E}\left[F(x_T)\right] - F(x_0) + \left(L - \frac{1}{2\eta}\right) \sum_{k=0}^{T-1} \|\hat{x}_{k+1} - x_k\|^2 \\
- \frac{\eta L^2}{\theta^2}\left(1 - \frac{1}{2n}\right) \sum_{i=1}^{n} \mathbb{E}\left\|x_T - \varphi_T^i\right\|^2 + \frac{\eta L^2}{\theta^2}\left(1 - \frac{1}{2n}\right) \sum_{i=1}^{n} \left\|x_0 - \varphi_0^i\right\|^2 .
\end{aligned}$$

Because $\varphi_0^i = x_0$ for all i, both of the terms on the second line are non-positive, so we can drop them from the inequality. Using the fact that $-F(x_T) \leq -F(x^*)$, our inequality simplifies to

$$-\left(L - \frac{1}{2\eta}\right) \sum_{k=0}^{T-1} \mathbb{E}\|\hat{x}_{k+1} - x_k\|^2 \leq F(x_0) - F(x^*).$$

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

$$\sum_{k=0}^{T-1} \mathbb{E}\|\mathcal{G}(x_k)\|^2 \leq \frac{F(x_0) - F(x^*)}{\eta^2\left(\frac{1}{2\eta} - L\right)}.$$

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

$$\mathbb{E}\|\mathcal{G}(x_\alpha)\|^2 \leq \frac{F(x_0) - F(x^*)}{\eta^2\left(\frac{1}{2\eta} - L\right)T}.$$

Choosing $\eta = \frac{\theta}{2Ln}$, which satisfies the bound in (E.3), gives the desired result.

Case 2. Suppose $\theta \geq 2$. Starting with (E.3), we apply the bound

$$\frac{\eta}{2}\left(1 - \frac{2}{\theta}\right)\left\|\nabla f(x_k) - \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\varphi_k^i)\right\|^2 \leq \frac{\eta L^2}{2n}\left(1 - \frac{2}{\theta}\right) \sum_{i=1}^{n} \left\|x_k - \varphi_k^i\right\|^2 .$$

This yields

$$\begin{aligned}
0 \leq -\mathbb{E}_k\left[F(x_{k+1})\right] + F(x_k) + \left(L - \frac{1}{2\eta}\right)\|\hat{x}_{k+1} - x_k\|^2 + \left(\frac{L}{2} - \frac{1}{2\eta}\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2 \\
+ \frac{\eta L^2}{2n}\left(1 - \frac{2}{\theta} + \frac{1}{\theta^2}\right) \sum_{i=1}^{n} \|x_k - \varphi_k^i\|^2.
\end{aligned}$$

Applying Lemma 14,

$$\begin{aligned}
0 \leq -\mathbb{E}_k\left[F(x_{k+1})\right] - F(x_k) + \left(L - \frac{1}{2\eta}\right)\|\hat{x}_{k+1} - x_k\|^2 \\
- \frac{c\eta L^2}{2}\left(1 - \frac{2}{\theta} + \frac{1}{\theta^2}\right)(1 - \delta) \sum_{i=1}^{n} \mathbb{E}_k\left\|x_{k+1} - \varphi_{k+1}^i\right\|^2 \\
+ \frac{c\eta L^2}{2}\left(1 - \frac{2}{\theta} + \frac{1}{\theta^2}\right)\left(1 - \frac{c-1}{cn}\right) \sum_{i=1}^{n} \left\|x_k - \varphi_k^i\right\|^2 \\
+ \left(\frac{cn\delta^{-1}\eta L^2}{2}\left(1 - \frac{2}{\theta} + \frac{1}{\theta^2}\right) + \frac{L}{2} - \frac{1}{2\eta}\right)\mathbb{E}_k\|x_{k+1} - x_k\|^2. \qquad (15)
\end{aligned}$$

As before, we set $\delta = \frac{c-1}{cn}$ and $c = 2$. The final term is non-positive as long as

$$\eta \leq \frac{1}{8Ln^2(\theta - 1)^2}\left(\sqrt{16n^2\theta^2(\theta^2 - 2\theta + 1) + \theta^4} - \theta^2\right)$$

With these parameter choices, we can drop the final term in (E.3). Summing the resulting inequality over the iterations $k = 0$ to $k = T - 1$ and following the same telescoping procedure as in Case 1, we are left with

$$-\left(L - \frac{1}{2\eta}\right) \sum_{k=1}^{T-1} \mathbb{E}\|\hat{x}_{k+1} - x_k\|^2 \leq F(x_0) - F(x^*).$$

Rewriting the term on the left in terms of $\mathcal{G}(x_\alpha)$ as in Case 1, we have the final bound

$$\mathbb{E}\|\mathcal{G}(x_\alpha)\|^2 \leq \frac{F(x_0) - F(x^*)}{\eta^2 \left(\frac{1}{2\eta} - L\right) T}.$$

Choosing $\eta = \frac{\theta}{2Ln(\theta-1)}$ proves the result.

□

F Proofs for B-SVRG

F.1 MSE bounds

Lemma 18 (Restatement of Theorem 2). *For $k \in \{ms, ms+1, \cdots, m(s+1)-1\}$, the MSE of the SVRG gradient estimator satisfies*

$$\mathbb{E}_k\|\widetilde{\nabla}_{\text{SVRG}} f(x_k) - \nabla f(x_k)\|^2 \leq \frac{L^2}{\theta^2}\|x_k - \varphi_s\|^2 + \left(1 - \frac{2}{\theta}\right)\|\nabla f(x_k) - \nabla f(\varphi_s)\|^2$$

Proof. This follows from Lemma 13 and the fact that $\varphi_k^i = \varphi_s$ for all i. □

Lemma 19. *For $k \in \{ms, ms+1, \cdots, m(s+1)-1\}$, the following inequality holds:*

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

Proof. Using the inequality $\|u - w\|^2 \leq (1+\delta)\|u - v\|^2 + (1+\delta^{-1})\|v - w\|^2$,

$$\begin{aligned}\|x_k - \varphi_s\|^2 &\leq (1+\delta)\|x_{k-1} - \varphi_s\|^2 + (1+\delta^{-1})\|x_k - x_{k-1}\|^2 \\ &\leq (1+\delta^{-1}) \sum_{\ell=ms}^{k} (1+\delta)^{k-\ell}\|x_{\ell+1} - x_\ell\|^2 \\ &\leq (1+\delta^{-1})(1+\delta)^m \sum_{\ell=ms}^{k} \|x_{\ell+1} - x_\ell\|^2\end{aligned}$$

The final inequality uses the estimate $(1+\delta)^{k-\ell} \leq (1+\delta)^m$. With $\delta = \frac{1}{m}$, summing this inequality from $k = ms$ to $k = m(s+1) - 1$ gives us

$$\begin{aligned}\sum_{k=ms}^{m(s+1)-1} \|x_k - \varphi_s\|^2 &\leq (m+1)\left(1 + \frac{1}{m}\right)^m \sum_{k=ms}^{m(s+1)-1} \sum_{\ell=ms}^{k} \|x_{\ell+1} - x_\ell\|^2 \\ &\leq m(m+1)\left(1 + \frac{1}{m}\right)^m \sum_{k=ms}^{m(s+1)-1} \|x_{\ell+1} - x_\ell\|^2 \\ &\leq 3m(m+1) \sum_{k=ms}^{m(s+1)-1} \|x_{k+1} - x_k\|^2.\end{aligned}$$

The final inequality uses the fact that $\left(1 + \frac{1}{m}\right)^m < \lim_{m\to\infty}\left(1 + \frac{1}{m}\right)^m = e < 3$, where e is Euler's constant. □

F.2 Convex

Theorem 20 (Restatement of Theorem 4, Part 1). *In Algorithm 2, set* $\eta = \frac{\theta}{4Lm(m+1)(\theta-1)+L(\theta+3\sqrt{2m(m+1)}}$ *for* $\theta \in [1,2)$ *and set* $\eta = \frac{\theta}{4Lm(m+1)(\theta-1)+3L\sqrt{2m(m+1)}(\theta-1)+L\theta}$ *for* $\theta \geq 2$. *After* S *epochs, Algorithm 2 produces an iterate satisfying*

$$\mathbb{E}\left[F(\overline{x}) - F(x^*)\right] \leq \begin{cases} \frac{2Lm(m+1)(\theta-1)+\frac{L}{2}(\theta+3\sqrt{2m(m+1)})}{mS\theta}\|x_0 - x^*\|^2 & \theta \in [1,2), \\ \frac{2Lm(m+1)(\theta-1)+\frac{3\sqrt{2}}{2}L\sqrt{m(m+1)}(\theta-1)+\frac{L\theta}{2}}{mS\theta}\|x_0 - x^*\|^2 & \theta \geq 2. \end{cases}$$

Proof. To begin, we consider the performance of Algorithm 2 over the single epoch s, so $k \in \{ms, ms+1, \cdots, m(s+1)-1$. The operator $\mathbb{E}_s$ denotes the expectation conditioned on the first $s-1$ epochs. From Lemma 12 with $\mu = 0$, we have

$$\begin{aligned} 0 \leq &\mathbb{E}_s\Bigg[-\eta F(x_{k+1}) + \eta F(x^*) + \frac{\eta}{2L\lambda}\|\widetilde{\nabla}_{\text{SVRG}} f(x_k) - \nabla f(x_k)\|^2 - \frac{1}{2}\|x_{k+1} - x^*\|^2 \\ &+ \frac{1}{2}\|x_k - x^*\|^2 + \eta\left(1 - \frac{1}{\theta}\right)\|x_k - \varphi_s\|^2 + \left(\frac{\eta L(\lambda+1)}{2} - \frac{1}{2}\right)\|x_{k+1} - x_k\|^2\Bigg], \end{aligned}$$

where we have used the fact that $\varphi_k^i = \varphi_s$ for all i. Using our bound on the MSE, we have

$$\begin{aligned} 0 \leq &\mathbb{E}_s\Bigg[-\eta F(x_{k+1}) + \eta F(x^*) + \frac{\eta}{2L\lambda}\left(1 - \frac{2}{\theta}\right)\|\nabla f(x_k) - \nabla f(\varphi_s)\|^2 \\ &- \frac{1}{2}\|x_{k+1} - x^*\|^2 + \frac{1}{2}\|x_k - x^*\|^2 + \left(\frac{\eta L(\lambda+1)}{2} - \frac{1}{2}\right)\|x_{k+1} - x_k\|^2 \\ &+ \frac{\eta L}{2}\left(1 - \frac{1}{\theta} + \frac{1}{\lambda\theta^2}\right)\|x_k - \varphi_s\|^2\Bigg]. \end{aligned} \tag{16}$$

As before, we consider two cases depending on θ.

Case 1. Suppose $\theta \in [1,2)$. In this case, $1 - \frac{2}{\theta} \leq 0$, so we can drop the second term in (F.2). Summing the resulting inequality from $k = ms$ to $k = m(s+1)-1$ gives

$$\begin{aligned} 0 \leq &\sum_{k=ms}^{m(s+1)-1} \mathbb{E}_s\left[-\eta F(x_{k+1}) + \eta F(x^*)\right] \\ &- \frac{1}{2}\mathbb{E}_s\|x_{m(s+1)} - x^*\|^2 + \frac{1}{2}\|x_{ms} - x^*\|^2 + \left(\frac{\eta L(\lambda+1)}{2} - \frac{1}{2}\right)\sum_{k=ms}^{m(s+1)-1}\mathbb{E}_s\|x_{k+1} - x_k\|^2 \\ &+ \frac{\eta L}{2}\left(1 - \frac{1}{\theta} + \frac{1}{\lambda\theta^2}\right)\sum_{k=ms}^{m(s+1)-1}\mathbb{E}_s\|x_k - \varphi_s\|^2. \end{aligned}$$

We use Lemma 19 the bound the final term. This gives

$$\begin{aligned} 0 \leq &\sum_{k=ms}^{m(s+1)-1} \mathbb{E}_s\left[-\eta F(x_{k+1}) + \eta F(x^*)\right] \\ &- \frac{1}{2}\mathbb{E}_s\|x_{m(s+1)} - x^*\|^2 + \frac{1}{2}\|x_{ms} - x^*\|^2 \\ &+ \left(\frac{\eta L(\lambda+1)}{2} + \frac{3m(m+1)\eta L}{2}\left(1 - \frac{1}{\theta} + \frac{1}{\lambda\theta^2}\right) - \frac{1}{2}\right)\sum_{k=ms}^{m(s+1)-1}\mathbb{E}_s\|x_{k+1} - x_k\|^2. \end{aligned}$$

We choose λ and η so that the final term is non-positive. To minimise this term over λ, we choose $\lambda = \frac{\sqrt{3m(m+1)}}{\theta}$. Following the step size used in our analysis of B-SAGA for simplicity, we choose

$$\eta = \frac{\theta}{4Lm(m+1)(\theta-1) + L(\theta + 3\sqrt{2m(m+1)})},$$

which is small enough to ensure that the terms in the round brackets are non-positive. (The optimal choice for η is larger only by a small constant factor.) This gives us the inequality

$$\begin{aligned} 0 \leq & \sum_{k=ms}^{m(s+1)-1} \mathbb{E}_s \left[-\eta F(x_{k+1}) + \eta F(x^*) \right] \\ & - \frac{1}{2} \mathbb{E}_s \| x_{m(s+1)} - x^* \|^2 + \frac{1}{2} \| x_{ms} - x^* \|^2 . \end{aligned}$$

Substituting the values $\varphi_{s+1} = x_{m(s+1)}$ and $\varphi_s = x_{ms}$, this becomes

$$\begin{aligned} 0 \leq & \sum_{k=ms}^{m(s+1)-1} \mathbb{E}_s \left[-\eta F(x_{k+1}) + \eta F(x^*) \right] \\ & - \frac{1}{2} \mathbb{E}_s \| \varphi_{s+1} - x^* \|^2 + \frac{1}{2} \| \varphi_s - x^* \|^2 . \end{aligned}$$

We can now chain this inequality and the conditional expectations over the epochs $s = 0$ to $s = S - 1$.

$$\sum_{k=0}^{mS-1} \mathbb{E} \left[\eta F(x_{k+1}) - \eta F(x^*) \right] \leq -\frac{1}{2} \mathbb{E} \| \varphi_S - x^* \|^2 + \frac{1}{2} \| x_0 - x^* \|^2 \leq \frac{1}{2} \| x_0 - x^* \|^2 .$$

Define $\overline{x} \overset{\text{def}}{=} \frac{1}{mS} \sum_{k=1}^{mS} x_k$. Using the convexity of F, we have shown

$$\eta m S \mathbb{E} \left[F(\overline{x}) - F(x^*) \right] \leq \eta \sum_{k=0}^{mS-1} \mathbb{E} \left[F(x_{k+1}) - F(x^*) \right] \leq \frac{1}{2} \| x_0 - x^* \|^2 .$$

This implies that

$$\mathbb{E} \left[F(\overline{x}) - F(x^*) \right] \leq \frac{1}{2 \eta m S} \| x_0 - x^* \|^2 ,$$

and substituting our choice for η completes the proof.

Case 2. We now consider the case $\theta \geq 2$. From equation (E.1), we have the inequality

$$\begin{aligned} 0 \leq & - \eta \mathbb{E}_s \Bigg[F(x_{k+1}) - F(x^*) - \frac{1}{2} \| x_{k+1} - x^* \|^2 + \frac{1}{2} \| x_k - x^* \|^2 \\ & + \left(\frac{\eta L (\lambda + 1)}{2} - \frac{1}{2} \right) \| x_{k+1} - x_k \|^2 + \frac{\eta L}{2} \left(1 + \frac{1}{\lambda} - \frac{1 + \frac{2}{\lambda}}{\theta} + \frac{1}{\lambda \theta^2} \right) \| x_k - \varphi_s \|^2 \Bigg] . \end{aligned}$$

Following the same procedure as in Case 1, we derive the bound

$$\begin{aligned} 0 \leq & \sum_{k=ms}^{m(s+1)-1} \mathbb{E}_s \left[-\eta F(x_{k+1}) + \eta F(x^*) \right] \\ & - \frac{1}{2} \mathbb{E}_s \| x_{m(s+1)} - x^* \|^2 + \frac{1}{2} \| x_{ms} - x^* \|^2 \\ & + \left(\frac{\eta L (\lambda + 1)}{2} + \frac{3m(m+1)\eta L}{2} \left(1 + \frac{1}{\lambda} - \frac{1 + \frac{2}{\lambda}}{\theta} + \frac{1}{\lambda \theta^2} \right) - \frac{1}{2} \right) \sum_{k=ms}^{m(s+1)-1} \mathbb{E}_s \| x_{k+1} - x_k \|^2 . \end{aligned}$$

To ensure the final term is non-positive, we set the parameters $\lambda = \frac{\sqrt{3m(m+1)}(\theta-1)}{\theta}$ and

$$\eta = \frac{\theta}{4Lm(m+1)(\theta - 1) + 3L\sqrt{2m(m+1)}(\theta - 1) + L\theta} .$$

With this value for η, the rest of the proof follows exactly as in Case 1.

□

F.3 Strongly convex

Theorem 21 (**Restatement of Theorem 4, Part 2**). *If g is μ-strongly convex in Algorithm 2, set $\eta = \frac{\theta}{5Lm(m+1)(\theta-1)+\frac{5L}{4}(\theta+3\sqrt{2m(m+1)})}$ for $\theta \in [1,2)$, and set $\eta = \frac{\theta}{5Lm(m+1)(\theta-1)+L\sqrt{2m(m+1)}(\theta-1)+L\theta}$. After S epochs, Algorithm 2 produces an iterate satisfying*

$$\mathbb{E}\|x_{mS}-x^*\|^2 \leq \begin{cases} \left(1+\frac{\mu\theta}{10Lm(m+1)(\theta-1)+\frac{5L}{2}(\theta+3\sqrt{2m(m+1)})}\right)^{-mS}\|x_0-x^*\|^2 & \text{for } \theta \in [1,2) \\ \left(1+\frac{\mu\theta}{10Lm(m+1)(\theta-1)+2L\sqrt{2m(m+1)}(\theta-1)+2L\theta}\right)^{-mS}\|x_0-x^*\|^2 & \text{for } \theta \geq 2. \end{cases}$$

Proof. We begin with inequality (F.2), but without setting $\mu = 0$.

$$\begin{aligned} 0 \leq \mathbb{E}_s\Bigg[& -\eta F(x_{k+1}) + \eta F(x^*) + \frac{\eta}{2L\lambda}\left(1-\frac{2}{\theta}\right)\|\nabla f(x_k) - \nabla f(\varphi_s)\|^2 \\ & - \frac{1+\mu\eta}{2}\|x_{k+1}-x^*\|^2 + \frac{1}{2}\|x_k - x^*\|^2 + \left(\frac{\eta L(\lambda+1)}{2} - \frac{1}{2}\right)\|x_{k+1}-x_k\|^2 \\ & + \frac{\eta L}{2}\left(1-\frac{1}{\theta}+\frac{1}{\lambda\theta^2}\right)\|x_k-\varphi_s\|^2\Bigg]. \end{aligned} \tag{17}$$

The term $-\eta(F(x_{k+1}) - F(x^*))$ is not useful to our analysis, so we drop it from the inequality. We proceed to analyse the same two cases as before.

Case 1. Suppose $\theta \in [1,2)$. In this case, $1 - \frac{2}{\theta} \leq 0$, so we can drop the second term in (F.3). We multiply this inequality by $(1+\mu\eta)^k$ and sum the result over the first epoch, from $k = 0$ to $k = m-1$.

$$\begin{aligned} 0 \leq & -\frac{(1+\mu\eta)^m}{2}\mathbb{E}_s\|x_m - x^*\|^2 + \frac{1}{2}\|x_0-x^*\|^2 + \left(\frac{\eta L(\lambda+1)}{2}-\frac{1}{2}\right)\sum_{k=0}^{m-1}(1+\mu\eta)^k\mathbb{E}_s\|x_{k+1}-x_k\|^2 \\ & + \frac{\eta L}{2}\left(1-\frac{1}{\theta}+\frac{1}{\lambda\theta^2}\right)\sum_{k=0}^{m-1}(1+\mu\eta)^k\mathbb{E}_s\|x_k-\varphi_s\|^2. \end{aligned}$$

With the choice

$$\eta = \frac{\theta}{5Lm(m+1)(\theta-1)+\frac{5L}{4}(\theta+3\sqrt{2m(m+1)})},$$

we can make the approximation $1+\mu\eta \leq 1 + \frac{\mu}{7Lm} \leq 1 + \frac{1}{7m}$. Hence, we can apply the estimate

$$\sum_{k=0}^{m-1}(1+\mu\eta)^k\mathbb{E}_s\|x_{k+1}-x_k\|^2 \leq (1+\mu\eta)^m\sum_{k=0}^{m-1}\mathbb{E}_s\|x_{k+1}-x_k\|^2 \leq \left(1+\frac{1}{7m}\right)^m\sum_{k=0}^{m-1}\mathbb{E}_s\|x_{k+1}-x_k\|^2.$$

We can bound the coefficient using the definition of Euler's number as in the proof of Lemma 18.

$$\left(1+\frac{1}{7m}\right)^m < \lim_{m\to\infty}\left(1+\frac{1}{7m}\right)^m = e^{1/7} < \frac{6}{5}.$$

This leaves us with the inequality

$$\begin{aligned} 0 \leq & -\frac{(1+\mu\eta)^m}{2}\mathbb{E}_s\|x_m - x^*\|^2 + \frac{1}{2}\|x_0-x^*\|^2 + \left(\frac{3\eta L(\lambda+1)}{5}-\frac{1}{2}\right)\sum_{k=0}^{m-1}\mathbb{E}_s\|x_{k+1}-x_k\|^2 \\ & + \frac{3\eta L}{5}\left(1-\frac{1}{\theta}+\frac{1}{\lambda\theta^2}\right)\sum_{k=0}^{m-1}\mathbb{E}_s\|x_k-\varphi_s\|^2. \end{aligned}$$

We now continue as in the proof of Theorem 20, using Lemma 19 the bound the final term. This gives

$$\begin{aligned}0 \leq &-\frac{(1+\mu\eta)^m}{2}\mathbb{E}_s\|x_m - x^*\|^2 + \frac{1}{2}\|x_0 - x^*\|^2 \\ &+ \left(\frac{3\eta L(\lambda+1)}{5} + \frac{9m(m+1)\eta L}{5}\left(1 - \frac{1}{\theta} + \frac{1}{\lambda\theta^2}\right) - \frac{1}{2}\right)\sum_{k=ms}^{m(s+1)-1}\mathbb{E}_s\|x_{k+1} - x_k\|^2.\end{aligned}$$

We choose λ and η so that the final term is non-positive. To minimise this term over λ, we choose $\lambda = \frac{\sqrt{3m(m+1)}}{\theta}$. These parameter choices ensure that the final term is non-positive. Hence, we have shown

$$\mathbb{E}_s\|x_m - x^*\|^2 \leq (1+\mu\eta)^{-m}\|x_0 - x^*\|^2.$$

Chaining this inequality and the expectations over epochs $s = 0$ to $s = S$ proves the desired result.

Case 2. Now suppose $\theta \geq 2$. As before, we apply the bound

$$\left(1 - \frac{2}{\theta}\right)\left\|\nabla f(x_k) - \frac{1}{n}\sum_{i=1}^n \nabla f_i(\varphi_k^i)\right\|^2 \leq L^2\left(1 - \frac{2}{\theta}\right)\sum_{i=1}^n \|x_k - \varphi_k^i\|^2$$

to inequality (F.3) to begin. Using the step size

$$\eta = \frac{\theta}{5Lm(m+1)(\theta-1) + L\sqrt{2m(m+1)}(\theta-1) + L\theta},$$

we are ensured that $1 + \mu\eta \leq 1 + \frac{1}{7n}$. Following the proof of Case 1, we have the inequality

$$\begin{aligned}0 \leq &-\frac{(1+\mu\eta)^m}{2}\mathbb{E}_s\|x_m - x^*\|^2 + \frac{1}{2}\|x_0 - x^*\|^2 \\ &+ \left(\frac{3\eta L(\lambda+1)}{5} + \frac{9m(m+1)\eta L}{5}\left(1 + \frac{1}{\lambda} - \frac{1+\frac{2}{\lambda}}{\theta} + \frac{1}{\lambda\theta^2}\right) - \frac{1}{2}\right)\sum_{k=ms}^{m(s+1)-1}\mathbb{E}_s\|x_{k+1} - x_k\|^2.\end{aligned}$$

We choose $\lambda = \frac{\sqrt{3m(m+1)}(\theta-1)}{\theta}$ to minimise the coefficient of the final term, and our choice for η ensures that this term is non-positive. Chaining the resulting inequality and the conditional expectations over epochs $s = 0$ to S finishes the proof. □

F.4 Non-convex

Theorem 22 (Restatement of Theorem 6). *In Algorithm 2, set $\eta = \frac{\theta}{2L\sqrt{m(m+1)}}$ for $\theta \leq 2$, and set $\eta = \frac{\theta}{2Ln(\theta-1)\sqrt{m(m+1)}}$ for $\theta \geq 2$. After T steps, Algorithm 2 achieves the following bound on the norm of the generalised gradient:*

$$\mathbb{E}\|\mathcal{G}_\eta(x_\alpha)\|^2 \leq \begin{cases} \frac{4L\sqrt{m(m+1)}(F(x_0)-F(x^*))}{\theta\left(1-\frac{\theta}{\sqrt{m(m+1)}}\right)T} & \text{for } 0 < \theta < 2, \\ \frac{4L\sqrt{m(m+1)}(\theta-1)(F(x_0)-F(x^*))}{\theta\left(1-\frac{\theta}{(\theta-1)\sqrt{m(m+1)}}\right)T} & \text{for } \theta \geq 2. \end{cases}$$

Proof. Define $\hat{x}$ as in (E.3). Following the proof of Theorem 17, we arrive at the inequality

$$\begin{aligned}\mathbb{E}_s\left[F(x_{k+1})\right] \leq \mathbb{E}_s\Bigg[&F(x_k) + \left(L - \frac{1}{2\eta}\right)\|\hat{x}_{k+1} - x_k\|^2 + \left(\frac{L}{2} - \frac{1}{2\eta}\right)\|x_{k+1} - x_k\|^2 \\ &+ \frac{\eta}{2}\|\nabla f(x_k) - \widetilde{\nabla}_{\text{SVRG}} f(x_k)\|^2\Bigg].\end{aligned}$$

Bounding the final term using Lemma 18, we have

$$0 \le \mathbb{E}_s \Bigg[- F(x_{k+1}) + F(x_k) + \left(L - \frac{1}{2\eta}\right) \|\hat{x}_{k+1} - x_k\|^2 + \left(\frac{L}{2} - \frac{1}{2\eta}\right) \|x_{k+1} - x_k\|^2 + \frac{\eta L^2}{2\theta^2}\|x_k - \varphi_s\|^2 + \frac{\eta}{2n}\left(1 - \frac{2}{\theta}\right)\sum_{i=1}^{n} \|\nabla f_i(x_k) - \nabla f_i(\varphi_s)\|^2 \Bigg]. \tag{18}$$

As before, we split our analysis into two cases depending on θ.

Case 1. Suppose $\theta \in (0, 2)$, so that $1 - \frac{2}{\theta} < 0$. We can simplify (F.4) to

$$0 \le \mathbb{E}_s \Bigg[- F(x_{k+1}) + F(x_k) + \left(L - \frac{1}{2\eta}\right) \|\hat{x}_{k+1} - x_k\|^2 + \left(\frac{L}{2} - \frac{1}{2\eta}\right) \|x_{k+1} - x_k\|^2 + \frac{\eta L^2}{2\theta^2}\|x_k - \varphi_s\|^2 \Bigg].$$

We bound the final term using Lemma 19.

$$0 \le \mathbb{E}_s \Bigg[- F(x_{k+1}) + F(x_k) + \left(L - \frac{1}{2\eta}\right) \|\hat{x}_{k+1} - x_k\|^2 + \left(\frac{L}{2} - \frac{1}{2\eta}\right) \|x_{k+1} - x_k\|^2 + \frac{\eta L^2 (1 + \delta^{-1})(1 + \delta)^m}{2\theta^2} \sum_{\ell = ms+1}^{k} \|x_\ell - x_{\ell-1}\|^2 \Bigg].$$

Summing this inequality over epoch s, which consists of iterations $k = ms$ to $k = m(s+1) - 1$, we have the inequality

$$0 \le \mathbb{E}_s \Bigg[- F(x_{m(s+1)}) + F(x_{ms}) + \left(L - \frac{1}{2\eta}\right) \sum_{k=ms}^{m(s+1)-1} \|\hat{x}_{k+1} - x_k\|^2 + \left(\frac{L}{2} - \frac{1}{2\eta}\right) \sum_{k=ms}^{m(s+1)-1} \|x_{k+1} - x_k\|^2 + \frac{\eta L^2 (1 + \delta^{-1})(1 + \delta)^m}{2\theta^2} \sum_{k=ms}^{m(s+1)-1} \sum_{\ell = ms+1}^{k} \|x_\ell - x_{\ell-1}\|^2 \Bigg].$$

We must choose δ and η so that the terms on the final line are non-positive. With $\delta = \frac{1}{m}$, we have that

$$\begin{aligned}(1+\delta)^m \sum_{k=ms}^{m(s+1)-1} \sum_{\ell = ms+1}^{k} \|x_\ell - x_{\ell-1}\|^2 &\le m \left(1 + \frac{1}{m}\right)^m \sum_{k=ms}^{m(s+1)-1} \|x_{k+1} - x_k\|^2 \\ &\le 3m \sum_{k=ms}^{m(s+1)-1} \|x_{k+1} - x_k\|^2.\end{aligned}$$

The final inequality uses the fact that $\left(1 + \frac{1}{m}\right)^m < \lim_{m \to \infty} \left(1 + \frac{1}{m}\right)^m = e < 3$, where e is Euler's number. With this bound, our inequality becomes

$$0 \le \mathbb{E}_s \Bigg[- F(x_{m(s+1)}) + F(x_{ms}) + \left(L - \frac{1}{2\eta}\right) \sum_{k=ms}^{m(s+1)-1} \|\hat{x}_{k+1} - x_k\|^2 + \left(\frac{L}{2} + \frac{\eta L^2 (1+m) 3m}{2\theta^2} - \frac{1}{2\eta}\right) \sum_{k=ms}^{m(s+1)-1} \|x_{k+1} - x_k\|^2 \Bigg].$$

The final term is non-positive as long as

$$\eta \le \frac{1}{6Lm(m+1)} \left(\sqrt{12 m \theta^2 (m+1) + \theta^4} - \theta^2\right).$$

For simplicity, we make the choice $\eta = \frac{\theta}{2L\sqrt{m(m+1)}}$. Substituting $x_{m(s+1)} = \varphi_{s+1}$ and $x_{ms} = \varphi_s$, we have

$$0 \leq \mathbb{E}_s \left[-F(\varphi_{s+1}) + F(\varphi_s) + \left(L - \frac{1}{2\eta} \right) \sum_{k=ms}^{m(s+1)-1} \|\hat{x}_{k+1} - x_k\|^2 \right].$$

Chaining this inequality and the expectations over the epochs $s = 0$ to $s = S - 1$,

$$0 \leq \mathbb{E}_s \left[-F(\varphi_S) + F(\varphi_0) + \left(L - \frac{1}{2\eta} \right) \sum_{k=0}^{mS-1} \|\hat{x}_{k+1} - x_k\|^2 \right].$$

Applying $-F(\varphi_S) \leq F(x^*)$, the definition of the generalised gradient, and the definition of x_α, we have our final inequality.

$$\mathbb{E}\|\mathcal{G}(x_\alpha)\|^2 \leq \frac{F(x_0) - F(x^*)}{\eta^2 \left(\frac{1}{2\eta} - L \right) T} = \frac{4L\sqrt{m(m+1)}(F(x_0) - F(x^*))}{\theta \left(1 + \frac{\theta}{\sqrt{m(m+1)}} \right) T},$$

where the final equality follows from our choice for η.

Case 2. For $\theta \geq 2$, we apply the bound

$$\left(1 - \frac{2}{\theta} \right) \left\| \nabla f(x_k) - \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\varphi_k^i) \right\|^2 \leq L^2 \left(1 - \frac{2}{\theta} \right) \sum_{i=1}^{n} \left\| x_k - \varphi_k^i \right\|^2$$

to inequality (F.4) just as in the convex case. This produces the inequality

$$\begin{aligned} 0 \leq \mathbb{E}_s \Bigg[& -F(x_{k+1}) + F(x_k) + \left(L - \frac{1}{2\eta} \right) \|\hat{x}_{k+1} - x_k\|^2 + \left(\frac{L}{2} - \frac{1}{2\eta} \right) \|x_{k+1} - x_k\|^2 \\ & + \frac{\eta L^2}{2} \left(1 - \frac{2}{\theta} + \frac{1}{\theta^2} \right) \|x_k - \varphi_s\|^2 \Bigg]. \end{aligned}$$

Following the same procedure as in Case 1, we arrive at the inequality

$$\begin{aligned} 0 \leq \mathbb{E}_s \Bigg[& -F(x_{m(s+1)}) + F(x_{ms}) + \left(L - \frac{1}{2\eta} \right) \sum_{k=ms}^{m(s+1)-1} \|\hat{x}_{k+1} - x_k\|^2 \\ & + \left(\frac{L}{2} + \frac{\eta L^2 (1+m) 3m}{2} \left(1 - \frac{2}{\theta} + \frac{1}{\theta^2} \right) - \frac{1}{2\eta} \right) \sum_{k=ms}^{m(s+1)-1} \|x_{k+1} - x_k\|^2 \Bigg]. \end{aligned}$$

The final term is non-positive for η satisfying

$$\eta \leq \frac{\sqrt{12m(m+1)\theta^2(\theta-1)^2 + \theta^4} - \theta^2}{6Lm(m+1)(\theta-1)^2}.$$

We make the particular choice $\eta = \frac{\theta}{2L(\theta-1)\sqrt{m(m+1)}}$. Applying the same telescoping procedure as in Case 1, we have

$$\begin{aligned} \mathbb{E}\|\mathcal{G}(x_\alpha)\|^2 &\leq \frac{F(x_0) - F(x^*)}{\eta^2 \left(\frac{1}{2\eta} - L \right) T} \\ &= \frac{4L(\theta-1)\sqrt{m(m+1)}(F(x_0) - F(x^*))}{\theta \left(1 + \frac{\theta}{(\theta-1)\sqrt{m(m+1)}} \right) T}. \end{aligned}$$

This completes the proof. □