Almost Tune-Free Variance Reduction

Bingcong Li* Lingda Wang† Georgios B. Giannakis*

** University of Minnesota - Twin Cities, Minneapolis, MN 55455, USA*
{lixx5599, georgios}@umn.edu

† University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
lingdaw2@illinois.edu

August 27, 2019

Abstract

The variance reduction class of algorithms including the representative ones, abbreviated as SVRG and SARAH, have well documented merits for empirical risk minimization tasks. However, they require grid search to optimally tune parameters (step size and the number of iterations per inner loop) for best performance. This work introduces ‘almost tune-free’ SVRG and SARAH schemes by equipping them with Barzilai-Borwein (BB) step sizes. To achieve the best performance, both i) averaging schemes; and, ii) the inner loop length are adjusted according to the BB step size. SVRG and SARAH are first reexamined through an ‘estimate sequence’ lens. Such analysis provides new averaging methods that tighten the convergence rates of both SVRG and SARAH theoretically, and improve their performance empirically when the step size is chosen large. Then a simple yet effective means of adjusting the number of iterations per inner loop is developed, which completes the tune-free variance reduction together with BB step sizes. Numerical tests corroborate the proposed methods.

1 Introduction

In this work, we deal with the frequently encountered empirical risk minimization (ERM) task expressed as

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) := \frac{1}{n}\sum_{i\in[n]} f_i(\mathbf{x}) \tag{1}$$

where $\mathbf{x}\in\mathbb{R}^d$ is the parameter vector to be learned from data; the set $[n] := \{1, 2, \ldots, n\}$ collects data indices; and, f_i is the loss function associated with datum i. Suppose that f is μ-strongly convex and has L-Lipchitz continuous gradient, with condition number denoted by $\kappa := L/\mu$. Throughout, $\mathbf{x}^*$ denotes the optimal solution of (1). The standard solver of (1) relies on *gradient descent* (GD), e.g. [Nesterov, 2004], which updates the parameter iterates via

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta\nabla f(\mathbf{x}_k)$$

where k is the iteration index and η the step size (or learning rate). For a strongly convex f, GD convergences linearly to $\mathbf{x}^*$; that is, $\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq (c_\kappa)^k\|\mathbf{x}_0 - \mathbf{x}^*\|^2$ for some κ-dependent constant $c_\kappa \in (0, 1)$ [Nesterov, 2004].

In the big data regime however, where n is huge, obtaining the gradient per iteration can be computationally prohibitive. To cope with this, the *stochastic gradient descent* (SGD) reduces the computational burden by drawing uniformly at random an index $i_k \in [n]$ per iteration k, and adopting $\nabla f_{i_k}(\mathbf{x}_k)$ as an unbiased estimator of $\nabla f(\mathbf{x}_k)$.

Albeit computationally lightweight with the simple update

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k\nabla f_{i_k}(\mathbf{x}_k)$$

the price paid is that SGD comes with sublinear convergence that is slower than GD [Robbins and Monro, 1951, Bottou et al., 2016]. It has been long recognized that the variance $\mathbb{E}[\|\nabla f_{i_t}(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\|^2]$ of the gradient estimator affects critically SGD’s convergence slowdown.

This naturally motivated gradient estimators with *reduced variance* compared with SGD's simple $\nabla f_{i_k}(\mathbf{x}_k)$. A gradient estimator with reduced variance can be obtained by capitalizing on the finite sum structure of (1). One idea is to judiciously evaluate a so-termed *snapshot gradient* $\nabla f(\mathbf{x}_s)$, and use it as an anchor of the stochastic draws in subsequent iterations. Members of the variance reduction family include schemes abbreviated as SDCA [Shalev-Shwartz and Zhang, 2013], SVRG [Johnson and Zhang, 2013], SAG [Roux et al., 2012], SAGA [Defazio et al., 2014], MISO [Mairal, 2013], SARAH [Nguyen et al., 2017], and their variants [Konecnỳ and Richtárik, 2013, Lei et al., 2017, Li et al., 2019, Kovalev et al., 2019]. Most of these algorithms rely on the update $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \mathbf{v}_k$, where η is a constant step size and $\mathbf{v}_k$ is an algorithm-specific gradient estimate that takes advantage of the snapshot gradient. In this work, SVRG and SARAH are of central interest because they are memory efficient compared with SAGA, and have no requirement for the duality arguments that SDCA entails. As elaborated later, we treat both SVRG as well as SARAH, with the latter allowing for bias in the gradient estimator. Variance reduction methods converge linearly when f is strongly convex. To fairly compare the complexity of (S)GD with that of variance reduction algorithms which combine snapshot gradients with the stochastic ones, we will rely on the incremental first-order oracle (IFO) [Agarwal and Bottou, 2015].

Definition 1. *An IFO takes f_i and $\mathbf{x} \in \mathbb{R}^d$ as input, and returns the (incremental) gradient $\nabla f_i(\mathbf{x})$.*

A desirable algorithm obtains an ϵ-accurate solution satisfying $\mathbb{E}[\|\nabla f(\mathbf{x})\|^2] \leq \epsilon$ or $\mathbb{E}[f(\mathbf{x}) - f(\mathbf{x}^*)] \leq \epsilon$ with minimal IFO complexity for a prescribed ϵ. IFO complexity for variance reduction alternatives such as SVRG and SARAH is $\mathcal{O}\big((n+\kappa)\ln\frac{1}{\epsilon}\big)$, a clear improvement over GD's complexity $\mathcal{O}\big(n\kappa \ln \frac{1}{\epsilon}\big)$. And when high accuracy (ϵ small) is desired, the complexity of variance reduction solvers is also lower than SGD's complexity of $\mathcal{O}\big(\frac{1}{\epsilon}\big)$. The merits of gradient estimators with reduced variance go beyond convexity [Reddi et al., 2016, Allen-Zhu and Hazan, 2016, Fang et al., 2018, Nguyen et al., 2019], but nonconvex ERM problems are out of the present work's scope.

Though theoretically appealing, SVRG and SARAH entail grid search to tune the step size, which is often painstakingly hard and time consuming. An automatically tuned step size for SVRG was introduced in [Barzilai and Borwein, 1988] (BB) and [Tan et al., 2016]. However, since both SVRG and SARAH have a double-loop structure, the inner loop length also requires tuning in addition to the step size. The BB step size was designed for SARAH in [Liu et al., Yang et al., 2019], but with extra tuneable parameters. In a nutshell, 'tune-free' variance reduction algorithms still have desired aspects to investigate and fulfill.

Along with auto-tuned BB step sizes, this paper establishes that in order to obtain 'tune-free' SVRG and SARAH schemes, one must: i) develop novel types of gradient averaging adaptive to the chosen step size; and, ii) adjust the inner loop length along with step size as well. Averaging in iterative solvers with reduced variance gradient estimators is effected by the means of choosing the starting point of the next outer loop [Johnson and Zhang, 2013, Tan et al., 2016, Nguyen et al., 2017, Li et al., 2019]. The types of averaging considered so far have been employed as tricks to simplify proofs, while in the algorithm itself only the last iteration is selected as starting point of the ensuing outer loop. However, we contend that different averaging results in different performance. And the best averaging depends on how large the step size is, which suggests that one should adjust the type of averaging in accordance with the step size. In addition to averaging, we argue that the choice of the inner loop length for BB-SVRG in [Tan et al., 2016] is too pessimistic. Addressing this with a simple modification leads to the desired 'almost tune-free' SVRG and SARAH solvers.

Our detailed contributions can be summarized as follows.

- We empirically argue that averaging is not merely a proof trick. It is prudent to adjust averaging in accordance with the step size chosen.
- SVRG and SARAH are analyzed using the notion of estimate sequence (ES). This prompts a novel averaging that tightens up convergence rate for SVRG, and further improves SARAH's convergence over existing works under certain conditions. Besides tighter rates, the novel analysis for SARAH broadens the analytical tool, estimate sequence, by endowing it with the ability to deal with SARAH's biased gradient estimator.
- We establish theoretical guarantees for BB-SVRG and BB-SARAH with different types of averaging, which broaden the range of selecting BB step sizes.
- Finally, we offer a principled design of the inner loop length and step size to obtain tune-free BB-SVRG and BB-SARAH. Numerical tests further corroborate the efficiency of the proposed algorithms.

Notation. Bold lowercase letters denote column vectors; $\mathbb{E}$ represents expectation; $\|\mathbf{x}\|$ stands for the ℓ_2-norm of $\mathbf{x}$; and $\langle \mathbf{x}, \mathbf{y} \rangle$ denotes the inner product of vectors $\mathbf{x}$ and $\mathbf{y}$.

2 Preliminaries

This section reviews the vanilla SVRG and SARAH starting with the basic assumptions on f.

2.1 Basic Assumptions

Assumption 1. *Each $f_i : \mathbb{R}^d \to \mathbb{R}$ has L-Lipchitz gradient; that is, $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

Assumption 2. *Each $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex.*

Assumption 3. *Function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex; that is, there exists $\mu > 0$, such that $f(\mathbf{x}) - f(\mathbf{y}) \geq \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

Assumption 4. *Each $f_i : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex, meaning there exists $\mu > 0$, so that $f_i(\mathbf{x}) - f_i(\mathbf{y}) \geq \langle \nabla f_i(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

Assumption 1 requires each loss function to be sufficiently smooth. One can certainly require smoothness of each individual loss function and refine Assumption 1 as f_i has L_i-Lipchitz gradient. Clearly $L = \max_i L_i$. By slightly modifying the algorithms, such a refined assumption can slightly tighten the bound in our analysis, and yield more desirable dependence of the IFO complexity on κ. However, since the extension is straightforward and along the lines of those appeared in [Xiao and Zhang, 2014, Kulunchakov and Mairal, 2019], we will keep using the simpler Assumption 1 for clarity. Assumption 3 only requires f to be strongly convex, which is weaker than Assumption 4. Assumptions 1 – 4 are all standard in variance reduction algorithms.

2.2 Recap of SVRG and SARAH

Algorithm 1 SVRG

```
Initialize: $\tilde{\mathbf{x}}^0$, $\eta$, $m$, $S$
for $s = 1, 2, \ldots, S$ do
    $\mathbf{x}_0^s = \tilde{\mathbf{x}}^{s-1}$
    $\mathbf{g}^s = \nabla f(\mathbf{x}_0^s)$
    for $k = 0, 1, \ldots, m-1$ do
        uniformly draw $i_k \in [n]$
        $\mathbf{v}_k^s = \nabla f_{i_k}(\mathbf{x}_k^s) - \nabla f_{i_k}(\mathbf{x}_0^s) + \mathbf{g}^s$
        $\mathbf{x}_{k+1}^s = \mathbf{x}_k^s - \eta \mathbf{v}_k^s$
    end for
    select $\tilde{\mathbf{x}}^s$ randomly from $\{\mathbf{x}_k^s\}_{k=0}^m$ following $\mathbf{p}^s$
end for
Output: $\tilde{\mathbf{x}}^S$
```

Algorithm 2 SARAH

```
Initialize: $\tilde{\mathbf{x}}^0$, $\eta$, $m$, $S$
for $s = 1, 2, \ldots, S$ do
    $\mathbf{x}_0^s = \tilde{\mathbf{x}}^{s-1}$, and $\mathbf{v}_0^s = \nabla f(\mathbf{x}_0^s)$
    $\mathbf{x}_1^s = \mathbf{x}_0^s - \eta \mathbf{v}_0^s$
    for $k = 1, 2, \ldots, m-1$ do
        uniformly draw $i_k \in [n]$
        $\mathbf{v}_k^s = \nabla f_{i_k}(\mathbf{x}_k^s) - \nabla f_{i_k}(\mathbf{x}_{k-1}^s) + \mathbf{v}_{k-1}^s$
        $\mathbf{x}_{k+1}^s = \mathbf{x}_k^s - \eta \mathbf{v}_k^s$
    end for
    select $\tilde{\mathbf{x}}^s$ randomly from $\{\mathbf{x}_k^s\}_{k=0}^m$ following $\mathbf{p}^s$
end for
Output: $\tilde{\mathbf{x}}^S$
```

The steps of SVRG and SARAH are listed in Algs. 1 and 2, respectively. Both employ a fine-grained reduced-variance gradient estimator per iteration. For SVRG, $\mathbf{v}_k^s$ is an unbiased estimator since $\mathbb{E}[\mathbf{v}_k^s|\mathcal{F}_{k-1}^s] = \nabla f(\mathbf{x}_k^s)$, where $\mathcal{F}_{k-1}^s := \sigma(\tilde{\mathbf{x}}^{s-1}, i_0, i_1, \ldots, i_{k-1})$; while SARAH adopts a biased $\mathbf{v}_k^s$, that is, $\mathbb{E}[\mathbf{v}_k^s|\mathcal{F}_{k-1}^s] = \nabla f(\mathbf{x}_k^s) - f(\mathbf{x}_{k-1}^s) + \mathbf{v}_{k-1}^s \neq \nabla f(\mathbf{x}_k^s)$. Unlike SGD though, the variance (equal to the mean-square error (MSE) for zero bias) of $\mathbf{v}_k^s$ in SVRG can be upper bounded by certain quantities that dictate the optimality gap (gradient norm square).

Lemma 1. *[Johnson and Zhang, 2013, Nguyen et al., 2017] The MSE of $\mathbf{v}_k^s$ in SVRG is bounded as follows*

$$\mathrm{SVRG}: \mathbb{E}\big[\|\nabla f(\mathbf{x}_k^s) - \mathbf{v}_k^s\|^2\big] \leq \mathbb{E}\big[\|\mathbf{v}_k^s\|^2\big] \leq 4L\mathbb{E}\big[f(\mathbf{x}_k^s) - f(\mathbf{x}^*)\big] + 4L\mathbb{E}\big[f(\mathbf{x}_0^s) - f(\mathbf{x}^*)\big].$$

The MSE of $\mathbf{v}_k^s$ in SARAH is also bounded as

$$\mathrm{SARAH}: \ \mathbb{E}\big[\|\nabla f(\mathbf{x}_k^s) - \mathbf{v}_k^s\|^2\big] \leq \frac{\eta L}{2 - \eta L}\bigg(\mathbb{E}\big[\|\nabla f(\mathbf{x}_0^s)\|^2\big] - \mathbb{E}\big[\|\mathbf{v}_k^s\|^2\big]\bigg).$$

Another upper bound on SVRG's gradient estimator is available [Kulunchakov and Mairal, 2019], but it is not suitable for our analysis. Intuitively, Lemma 1 suggests that if SVRG or SARAH converges, the MSE of their gradient estimator also approaches zero.

At the end of each inner loop, the starting point of the next outer loop is randomly selected among $\{\mathbf{x}_k^s\}_{k=0}^m$ according to a probability mass function (pmf) vector $\mathbf{p}^s \in \Delta_{m+1}$, where $\Delta_{m+1} := \{\mathbf{p} \in \mathbb{R}_+^{m+1} | \langle \mathbf{1}, \mathbf{p} \rangle = 1\}$. We term $\mathbf{p}^s$ the *averaging weight vector*, and let p_j^s denote the jth entry of $\mathbf{p}^s$. Leveraging the MSE bounds in Lemma 1 and choosing a proper averaging vector, SVRG and SARAH iterates for strongly convex problems can be proved to be linearly convergent [Johnson and Zhang, 2013, Tan et al., 2016, Nguyen et al., 2017, Li et al., 2019].

To prove SVRG convergence, two types of averaging exist.

- **U-Avg (SVRG)** [Johnson and Zhang, 2013]: vector $\mathbf{p}^s$ is chosen as the pmf of an (almost) uniform distribution; that is, $p_m^s = 0$, and $p_k^s = 1/m$ for $k = \{0, 1, \ldots, m-1\}$. Under Assumptions 1 – 3, the choice of $\eta = \mathcal{O}(1/L)$ and $m = \mathcal{O}(\kappa)$ ensures that SVRG iterates converge linearly.[1]

- **L-Avg (SVRG)** [Tan et al., 2016]: Only the last iteration is used for averaging by setting $\tilde{\mathbf{x}}^s = \mathbf{x}_m^s$; or equivalently, by setting $p_m^s = 1$, and $p_k^s = 0, \forall k \neq m$. Under Assumptions 1 – 3, linear convergence is ensured by choosing $\eta = \mathcal{O}(1/(L\kappa))$ and $m = \mathcal{O}(\kappa^2)$.

To guarantee linear convergence, SVRG with L-Avg must adopt a much smaller η and larger m compared with U-Avg. L-Avg with such a small step size leads to IFO complexity $\mathcal{O}\big((n + \kappa^2) \ln \frac{1}{\epsilon}\big)$ that has worse dependence on κ.

For SARAH, there are also two averaging options.

- **U-Avg (SARAH)** [Nguyen et al., 2017]: here $\mathbf{p}^\mathbf{s}$ is selected to have entries $p_m^s = 0$, and $p_k^s = 1/m$, for $k = \{0, 1, \ldots, m-1\}$. Linear convergence is guaranteed with complexity $\mathcal{O}\big((n + \kappa) \ln \frac{1}{\epsilon}\big)$ under Assumptions 1 – 3 so long as one selects $\eta = \mathcal{O}(1/L)$ and $m = \mathcal{O}(\kappa)$.

- **L-Avg (SARAH)** [Li et al., 2019][2]: here $\mathbf{p}^s$ is chosen with entries $p_{m-1}^s = 1$ and $p_k^s = 0, \forall k \neq m-1$. Under Assumptions 1 – 3 and with $\eta = \mathcal{O}(1/L)$ as well as $m = \mathcal{O}(\kappa^2)$, linear convergence is guaranteed at IFO complexity of $\mathcal{O}\big((n + \kappa^2) \ln \frac{1}{\epsilon}\big)$. When both Assumptions 1 and 4 hold, setting $\eta = \mathcal{O}(1/L)$ and $m = \mathcal{O}(\kappa)$ results in linear convergence along with a reduced complexity of order $\mathcal{O}\big((n + \kappa) \ln \frac{1}{\epsilon}\big)$.

U-Avg (for both SVRG and SARAH) is usually employed as a 'proof-trick' to carry out convergence analysis, while L-Avg is implemented most of the times. However, we will argue in the next section that with U-Avg adapted to the step size choice it is possible to improve performance. Although U-Avg appears at first glance to waste updates, a simple trick in the implementation can also fix this issue.

Implementation of averaging. With weighted averaging effected by $\mathbf{p}^s$, rather than updating m times and then choosing $\tilde{\mathbf{x}}^s$ according to Line 10 of SVRG or SARAH, one can generate a random integer $M^s \in \{0, 1, \ldots, m\}$ according to $\mathbf{p}^s$. Having available $\mathbf{x}_{M^s}^s$, it is possible to start the next inner loop immediately.

3 Estimate Sequence for SVRG and SARAH

In this section, SVRG and SARAH are reexamined through the lens of an 'estimate sequence' (ES), a tool that has been used for analyzing momentum schemes [Nesterov, 2004]; see also [Nitanda, 2014, Lin et al., 2015, Kulunchakov and Mairal, 2019]. By permeating the benefits of ES to SVRG and SARAH, we will enable novel means of averaging. The novel averaging for SVRG will considerably tighten its analytical convergence rate; while for SARAH it will improve its convergence rate when m is chosen large enough. In addition, since existing ES analysis relies heavily on the unbiasedness of $\mathbf{v}_k^s$, our advances here will endow the ES tool with the ability to deal with biased gradient estimators, thus making it suitable also for SARAH.

[1] For simplicity and clarity of exposition we only highlight the order of η and m, and hide other constants in big-$\mathcal{O}$ notation. Detailed choices can be found in the corresponding references.

[2] There is another version of L-Avg for SARAH [Liu et al.], but convergence claims require undesirably small step sizes $\eta = \mathcal{O}(\mu/L^2)$. This is why we focus on the L-Avg in [Li et al., 2019].

3.1 Estimate Sequence

Since in this section we will focus on a specific inner loop indexed by s, we drop superscript s for brevity. For example, $\mathbf{x}_k^s$ and $\mathbf{v}_k^s$ are written as $\mathbf{x}_k$ and $\mathbf{v}_k$, respectively.

Associated with our ERM objective f and a particular point $\mathbf{x}_0$, consider the series of quadratic functions $\{\Phi_k(\mathbf{x})\}_{k=0}^m$ that comprise what we term ES, with the first one given by

$$\Phi_0(\mathbf{x}) = \Phi_0^* + \frac{\mu_0}{2}\|\mathbf{x} - \mathbf{x}_0\|^2 \tag{3a}$$

and the rest defined recursively as

$$\Phi_k(\mathbf{x}) =(1-\delta_k)\Phi_{k-1}(\mathbf{x}) + \delta_k\Big[f(\mathbf{x}_{k-1}) + \langle \mathbf{v}_{k-1}, \mathbf{x} - \mathbf{x}_{k-1}\rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}_{k-1}\|^2\Big]$$

where $\mathbf{v}_{k-1}$ is the gradient estimate in SVRG or SARAH; while Φ_0^*, μ_0, μ, and δ_k are some constants to be specified later. The design is similar to that of [Kulunchakov and Mairal, 2019], but the ES here is constructed per inner loop. In addition, here we will overcome the challenge of analyzing SARAH's biased gradient estimator $\mathbf{v}_k$.

Upon defining $\Phi_k^* := \min_{\mathbf{x}} \Phi_k(\mathbf{x})$, the key properties of the sequence $\{\Phi_k(\mathbf{x})\}_{k=0}^m$ are collected in the next lemma.

Lemma 2. *For $\{\Phi_k(\mathbf{x})\}_{k=0}^m$ as in* (3)*, it holds that:*
i) $\Phi_0(\mathbf{x})$ is μ_0-strongly convex, and $\Phi_k(\mathbf{x})$ is μ_k-strongly convex with $\mu_k = (1-\delta_k)\mu_{k-1} + \delta_k\mu$;
ii) $\mathbf{x}_k$ minimizes $\Phi_k(\mathbf{x})$ if $\delta_k = \eta\mu_k$; and
iii) $\Phi_k^* = (1-\delta_k)\Phi_{k-1}^* + \delta_k f(\mathbf{x}_{k-1}) - \frac{\mu_k\eta^2}{2}\|\mathbf{v}_{k-1}\|^2$.

Lemma 2 holds for both SVRG and SARAH. To better understand the role of ES, it is instructive to use an example.

Example. With $\Phi_0^* = f(\mathbf{x}_0)$, $\mu_0 = \mu$ and $\delta_k = \mu_k\eta$ for SVRG, it holds that $\mu_k = \mu, \forall k$, and $\delta_k = \mu\eta, \forall k$. If for convenience we let $\delta := \mu\eta$, we show in Appendix A.2 that

$$\mathbb{E}\big[\Phi_k(\mathbf{x})\big] \le (1-\delta)^k\big[\Phi_0(\mathbf{x}) - f(\mathbf{x}^*)\big] + f(\mathbf{x}). \tag{4}$$

As $k \to \infty$, one has $(1-\delta)^k \to 0$, and hence $\Phi_k(\mathbf{x})$ approaches in expectation a lower bound of $f(\mathbf{x})$.

Now, we are ready to view SVRG and SARAH through the lens of $\{\Phi_k(\mathbf{x})\}_{k=0}^m$.

3.2 ES for SVRG

The major byproduct offered by ES is a novel averaging vector $\mathbf{p}^s$, which can improve the convergence of SVRG.

Theorem 1. *Under Assumptions 1 – 3, consider the ES in* (3) *with $\mu_0 = \mu$, $\delta_k = \mu_k\eta$, and $\Phi_0^* = f(\mathbf{x}_0)$. Choose $\eta < 1/(4L)$, and m large enough such that*

$$\lambda^{\textit{SVRG}} := \frac{1}{1-(1-\mu\eta)^{m-1}}\bigg[\frac{(1-\mu\eta)^m}{1-2\eta L} + \frac{2\mu L\eta^2(1-\mu\eta)^{m-1}}{1-2L\eta} + \frac{2L\eta}{1-2L\eta}\bigg] < 1.$$

Let $p_0^s = p_m^s = 0$, and $p_k^s = (1-\mu\eta)^{m-k-1}/q$ for $k = 1, 2, \ldots, m-1$, where $q = [1-(1-\mu\eta)^{m-1}]/(\mu\eta)$. It then holds for SVRG with this weighted averaging (W-Avg) that

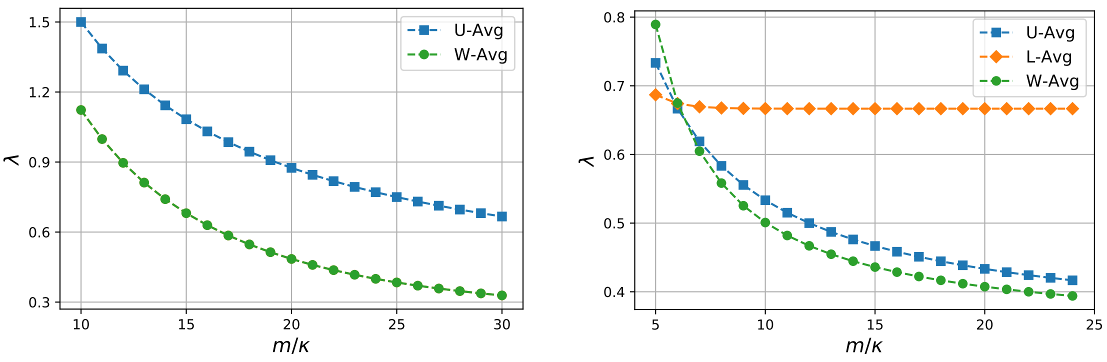
$$\mathbb{E}\big[f(\tilde{\mathbf{x}}^s) - f(\mathbf{x}^*)\big] \le \lambda^{\textit{SVRG}}\mathbb{E}\big[f(\tilde{\mathbf{x}}^{s-1}) - f(\mathbf{x}^*)\big].$$

Comparing the W-Avg in Theorem 1 against U-Avg and L-Avg we saw in Section 2.2, the upshot of W-Avg is a much tighter convergence rate; see also Fig. 1(a) comparing the theoretical convergence rates of different types of averaging. For provable convergence of SVRG with L-Avg, η and m should be chosen different from those in U-Avg and W-Avg. For this reason, L-Avg is not plotted in Fig. 1(a). As one can see that when choosing $\eta = \mathcal{O}(1/L)$, the dominate terms of the convergence rate for W-Avg is $\mathcal{O}\big(\frac{(1-1/\kappa)^m}{1-2L\eta} + \frac{2L\eta}{1-2L\eta}\big)$, while $\mathcal{O}\big(\frac{\kappa}{m(1-2L\eta)} + \frac{2L\eta}{1-2L\eta}\big)$ for U-Avg [Johnson and Zhang, 2013]. Clearly, $(1-1/\kappa)^m$ in the convergence rate of W-Avg can be much smaller than the $\frac{\kappa}{m}$ in that of U-Avg.

Next, we assess the IFO complexity of SVRG with W-Avg.

Corollary 1. *Choosing $m = \mathcal{O}(\kappa)$ and other parameters as in Theorem 1, the complexity of SVRG with W-Avg to find $\tilde{\mathbf{x}}^s$ satisfying $\mathbb{E}\big[f(\tilde{\mathbf{x}}^s) - f(\mathbf{x}^*)\big] \le \epsilon$, is $\mathcal{O}\big((n+\kappa)\ln\frac{1}{\epsilon}\big)$.*

Note that similar to U-Avg, W-Avg incurs lower IFO complexity compared with L-Avg in [Tan et al., 2016].



(a) SVRG

(b) SARAH

Figure 1: A comparison of the analytical convergence rate for SVRG and SARAH. In both figures we set $\kappa = 10^5$ with $L = 1$, $\mu = 10^{-5}$, and the step sizes are selected as: (a) SVRG with $\eta = 0.1/L$; and (b) SARAH with $\eta = 0.5/L$.

3.3 ES for SARAH

SARAH is challenging to analyze due to the bias present in the estimator $\mathbf{v}_k$, which makes the ES-based treatment of SARAH fundamentally different from that of SVRG. To see this, it is useful to start with the following lemma.

Lemma 3. *For any deterministic* $\mathbf{x}$*, it holds in SARAH that*

$$\mathbb{E}\big[\langle \mathbf{v}_k - \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k\rangle\big] = \frac{\eta}{2}\sum_{\tau=0}^{k-1}\mathbb{E}\Big[\|\mathbf{v}_\tau - \nabla f(\mathbf{x}_\tau)\|^2 + \|\mathbf{v}_\tau\|^2 - \|\nabla f(\mathbf{x}_\tau)\|^2\Big].$$

No assumption is needed for Lemma 3, which reveals the main difference in the ES-based argument for SARAH, namely that $\mathbb{E}\big[\langle \mathbf{v}_k - \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k\rangle\big] \neq 0$, while the same inner product for SVRG equals to 0 in expectation. Reflecting back to (4), the consequence of having a non-zero $\mathbb{E}\big[\langle \mathbf{v}_k - \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k\rangle\big]$ is that $\mathbb{E}[\Phi_k(\mathbf{x})]$ is no longer a lower bound of $f(\mathbf{x})$ as $k \to \infty$; thus,

$$\mathbb{E}\big[\Phi_k(\mathbf{x})\big] \leq (1-\delta)^k\big[\Phi_0(\mathbf{x}) - f(\mathbf{x})\big] + f(\mathbf{x}) + C \tag{5}$$

where C is a non-zero term that is not present in (4) when applied to SVRG; see detailed derivations in Appendix A.2.

Interestingly, upon capitalizing on properties of $\mathbf{v}_k$, the ensuing theorem establishes linear convergence for SARAH with a proper W-Avg vector $\mathbf{p}^s$.

Theorem 2. *Under Assumptions 1 and 4, consider the ES in (3) with* $\mu_0 = \mu$*,* $\delta_k = \mu_k\eta, \forall k$*, and* $\Phi_0^* = f(\mathbf{x}_0)$*. With* $\delta := \mu\eta$*, select* $\eta < 1/L$ *and* m *large enough, so that*

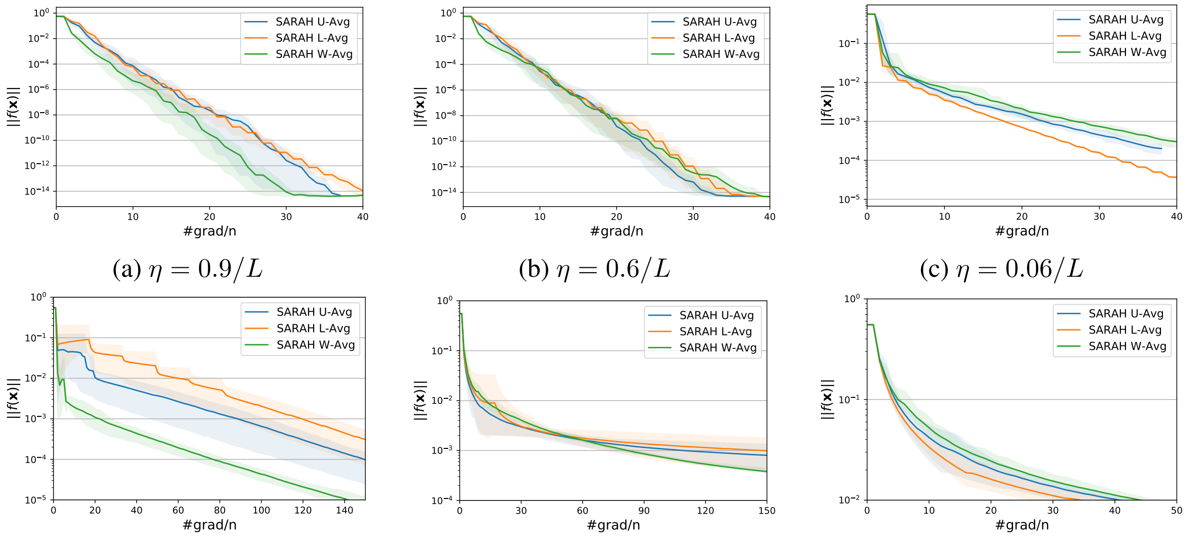
$$\lambda^{\textit{SARAH}} := \Big[(1-\delta)^m - \Big(1-\frac{2\eta L}{1+\kappa}\Big)^m\Big]\frac{L+\mu}{c(L-\mu)} + \frac{(1-\delta)^m}{c\delta} + \frac{\eta L(m-1)}{c(2-\eta L)} + \frac{2-2\eta L}{2-\eta L}\,\frac{1+\kappa}{2c\eta L} < 1$$

where $c = m - \frac{1}{\delta} + \frac{(1-\delta)^m}{\delta}$*. Setting* $p_k = (1-(1-\delta)^{m-k-1})/c, \forall k = 0, 1, \ldots, m-2$*, and* $p_{m-1} = p_m = 0$*, SARAH with this W-Avg satisfy*

$$\mathbb{E}\big[\|\nabla f(\tilde{\mathbf{x}}^s)\|^2\big] \leq \lambda^{\textit{SARAH}}\mathbb{E}\big[\|\nabla f(\tilde{\mathbf{x}}^{s-1})\|^2\big].$$

The expression of λ^{SARAH} is complicated because we want the upper bound of the convergence rate to be as tight as possible. To demonstrate this with an example, choosing $\eta = 1/(2L)$ and $m = 5\kappa$, we have $\lambda^{\text{SARAH}} \approx 0.8$. Fig. 1(b) compares SARAH with W-Avg versus SARAH with U-Avg and L-Avg. The advantage of W-Avg is more pronounced as m is chosen larger.

As far as IFO complexity of SARAH with W-Avg, it is comparable with that of L-Avg or U-Avg, as asserted next.



(a) $\eta = 0.9/L$ (b) $\eta = 0.6/L$ (c) $\eta = 0.06/L$

(d) $\eta = 0.1/L$ (e) $\eta = 0.01/L$ (f) $\eta = 0.005/L$

Figure 3: Comparing SARAH with different types of averaging on datasets *w7a* and *diabetes* ($\mu = 0.005$ and $m = 5\kappa$ is all tests).

Corollary 2. *Choosing $m = \mathcal{O}(\kappa)$ and other parameters as in Theorem 2, the complexity of SARAH with W-Avg to find $\tilde{\mathbf{x}}^s$ satisfying $\mathbb{E}[\|\nabla f(\tilde{\mathbf{x}}^s)\|^2] \leq \epsilon$, is $\mathcal{O}\big((n+\kappa)\ln\frac{1}{\epsilon}\big)$.*

A few remarks are now in order on our analytical findings: i) most existing ES-based proofs use $\mathbb{E}[f(\tilde{\mathbf{x}}^s) - f(\mathbf{x}^*)]$ as optimality metric, while Theorem 2 and Corollary 2 rely on $\mathbb{E}[\|\nabla f(\tilde{\mathbf{x}}^s)\|^2]$; ii) the analysis method still holds when Assumption 4 is weakened to Assumption 3, at the price of having worse κ-dependence of the IFO complexity; that is, $\mathcal{O}\big((n+\kappa^2)\ln\frac{1}{\epsilon}\big)$, which if of the same order as L-Avg under Assumptions 1 – 3 [Li et al., 2019, Liu et al.].

3.4 Averaging is More Than A ‘Proof Trick’

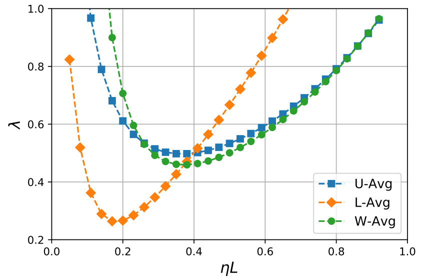


Figure 2: SARAH’s analytical convergence with different averaging options ($\kappa = 10^5$, $L = 1$, $\mu = 10^{-5}$, and fixed $m = 10\kappa$).

Existing forms of averaging such as U-Avg and W-Avg, are typically employed as ‘proof tricks’ for simplifying the theoretical analysis. In this subsection, we contend that averaging can distinctly affect performance, and should be adapted to the step size. Throughout this subsection we will consider SARAH with $\eta = \mathcal{O}(1/L)$ and $m = \mathcal{O}(\kappa)$ because this parameter selection guarantees convergence regardless of the averaging employed. For SVRG to converge on the other hand, the step size has to be chosen differently when replacing L-avg with W-Avg or U-Avg

We will first look at the convergence rate of SARAH across the different averaging options. Fixing m and changing η, the theoretical convergence rate is plotted in Fig. 2. It is seen that with smaller step sizes, L-Avg enjoys faster convergence, while larger step sizes tend to favor W-Avg and U-Avg instead.

Next, we will demonstrate empirically that the type of averaging indeed matters. Consider binary classification using the regularized logistic loss function

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i\in[n]} \ln\big[1+\exp(-b_i\langle\mathbf{a}_i,\mathbf{x}\rangle)\big] + \frac{\mu}{2}\|\mathbf{x}\|^2 \tag{6}$$

where $(\mathbf{a}_i, b_i)$ is the (feature, label) pair of datum i. Clearly, (6) is an instance of the cost in (1) with $f_i(\mathbf{x}) = \ln\big[1+\exp(-b_i\langle\mathbf{a}_i,\mathbf{x}\rangle)\big] + \frac{\mu}{2}\|\mathbf{x}\|^2$; and it can be readily verified that Assumptions 1 and 4 are also satisfied in this case.

We tested the performance of SARAH with L-Avg, U-Avg and W-Avg with fixed m but different step size choices on the datasets *w7a* and *diabetes*; see also Appendix D.1 for additional tests with dataset *a9a*. Fig. 3(a) shows that for a large step size $\eta = 0.9/L$, W-Avg outperforms U-Avg as well as L-Avg on *w7a*. For a medium step size $\eta = 0.6/L$, W-Avg and L-Avg perform comparably, while both are outperformed by U-Avg. When η is chosen small, L-Avg is clearly the winner. In short, the performance of averaging options varies with the step size. This is intuitively reasonable because: i) the MSE of $\mathbf{v}_k$ scales with η (cf. Lemma 1); and ii) it tends to increase with k as $\mathbb{E}[\|\mathbf{v}_k\|^2]$ decreases linearly (see Lemma 5 in Appendix B.2, and the MSE bound in Lemma 1). As a result, when both η and k are large, the MSE of $\mathbf{v}_k$ tends to be large too. Iterates with gradient estimators having high MSE can jeopardize the convergence. This explains the inferior performance of L-Avg in Figs. 3(a) and 3(b). On the other hand, when η is chosen small, the MSE tends to be small as well; hence, working with L-Avg does not compromise convergence, while in expectation W-Avg and U-Avg compute full gradient more frequently than L-Avg. These two reasons explain the improved performance of L-Avg in Fig. 3(c).

Such a performance difference among types of averaging suggests that one should also account for the averaging type to optimize performance when tuning the step sizes.

4 Tune-free Variance Reduction

Aiming to develop 'tune-free' SVRG and SARAH, we will first adopt the Barzilai-Borwein (BB) scheme to obtain suitable step sizes automatically [Tan et al., 2016]. In a nutshell, BB monitors progress of previous outer loops, and chooses the step size of outer loop s accordingly via

$$\eta^s = \frac{1}{\theta_\kappa} \frac{\|\tilde{\mathbf{x}}^{s-1} - \tilde{\mathbf{x}}^{s-2}\|^2}{\big\langle \tilde{\mathbf{x}}^{s-1} - \tilde{\mathbf{x}}^{s-2}, \nabla f(\tilde{\mathbf{x}}^{s-1}) - \nabla f(\tilde{\mathbf{x}}^{s-2}) \big\rangle} \tag{7}$$

where θ_κ is a κ-dependent parameter to be specified later. Note that $\nabla f(\tilde{\mathbf{x}}^{s-1})$ and $\nabla f(\tilde{\mathbf{x}}^{s-2})$ are computed at the outer loops s and $s-1$, respectively; hence, the implementation overhead of BB step sizes only includes almost negligible memory overhead to store $\tilde{\mathbf{x}}^{s-2}$ and $\nabla f(\tilde{\mathbf{x}}^{s-2})$.

BB step sizes for SVRG with L-Avg have relied on $\theta_\kappa = m = \mathcal{O}(\kappa^2)$ [Tan et al., 2016]. Such a choice of parameters offers provable convergence at complexity $\mathcal{O}\big((n+\kappa^2)\ln\frac{1}{\epsilon}\big)$, but has not been effective in our simulations for two reasons: i) step size η^s depends on m, which means that tuning is still required for step sizes; and, ii) the optimal m of $\mathcal{O}(\kappa)$ with best empirical performance significantly deviates from the theoretically suggested $\mathcal{O}(\kappa^2)$; see also Fig. 4. This prompted us to design more practical BB-step sizes. Besides step size, another important parameter that requires tuning is m. How to choose m with minimal tuning is also of major practical value.

4.1 Adapting Type of Averaging with BB Step Sizes

We start with a fixed choice of m to theoretically investigate the BB step sizes under different types of averaging. The final 'tune-free' implementation of SVRG and SARAH will rely on the analysis of this subsection.

One can verify that the BB step size in (7) lies in the following interval (see Appendix C for a proof)

$$\frac{1}{\theta_\kappa L} \leq \eta^s \leq \frac{1}{\theta_\kappa \mu}. \tag{8}$$

The choices of θ_κ and m in [Tan et al., 2016] result in a small upper bound on η^s, namely of $\mathcal{O}(1/(\kappa L))$. Such a step size is too small, and leads to the worst performance depicted in Fig. 4. As the upper bound on η^s is κ times larger than its lower bound, we deduce from (8) that the BB step size can change over a wide range. This suggests that averaging should be designed along with BB step sizes, as we assert next starting with BB-SVRG.

Proposition 1. ***(BB-SVRG)*** *Under Assumptions 1 – 3, if we choose $m = \mathcal{O}(\kappa^2)$ and $\theta_\kappa = \mathcal{O}(\kappa)$ (but with $\theta_\kappa > 4\kappa$), then BB-SVRG with U-Avg and W-avg can find $\tilde{\mathbf{x}}^s$ with $\mathbb{E}\big[f(\tilde{\mathbf{x}}^s) - f(\mathbf{x}^*)\big] \leq \epsilon$ using $\mathcal{O}\big((n+\kappa^2)\ln\frac{1}{\epsilon}\big)$ IFO calls.*

Different from BB-SVRG, the ensuing result asserts that for BB-SARAH, W-Avg, U-Avg and L-Avg have identical order of IFO complexity.

Proposition 2. ***(BB-SARAH)*** *Under Assumptions 1 and 4, if we choose $m = \mathcal{O}(\kappa^2)$ and $\theta_\kappa = \mathcal{O}(\kappa)$, then BB-SARAH finds a solution with $\mathbb{E}\big[\|\nabla f(\tilde{\mathbf{x}}^s)\|^2\big] \leq \epsilon$ using $\mathcal{O}\big((n+\kappa^2)\ln\frac{1}{\epsilon}\big)$ IFO calls, when one of these conditions holds: i) either U-Avg with $\theta_\kappa > \kappa$; or ii) L-Avg with $\theta_\kappa > 3/2\kappa$; or, iii) W-Avg with $\theta_\kappa > \kappa$.*

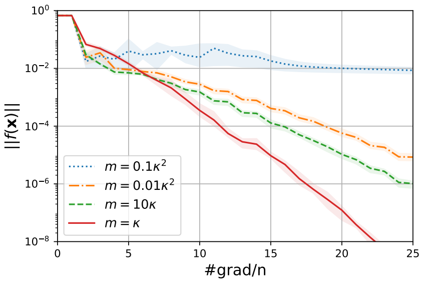


Figure 4: Performance of BB-SVRG [Tan et al., 2016] under different choices of m on dataset *a9a* ($\kappa = 1,388$ and $\theta_\kappa = m$).

In Propositions 1 and 2, choosing $\theta_\kappa = \mathcal{O}(\kappa)$ guarantees that $\mathcal{O}(1/L)$ upper bounds the BB step size in (8). This upper bound is the same step size order used in plain-vanilla SVRG and SARAH. The price paid for having automatically tuned step sizes is a worse dependence of the IFO complexity on κ, compared with the bounds in Corollaries 1 and 2. The cause of the worse dependence on κ is that one has to bear with small step sizes, as η^s approaches its lower bound in (8). However, such an automatic tuning of the step size comes almost as a "free lunch" when problem (1) is well conditioned, or, in the big data regime, e.g., $\kappa^2 \approx n$ or $\kappa^2 \ll n$, since the dominant term in IFO complexity is $\mathcal{O}(n \ln \frac{1}{\epsilon})$ for both SVRG and BB-SVRG. On the other hand, it is prudent to stress that with $\kappa^2 \gg n$, the BB step sizes slow down convergence.

Having theoretically guaranteed convergence with different types of averaging manifests itself to improved performance of BB-SVRRG and BB-SARAH. The general guideline is that when the step size is selected large enough, meaning $\eta = \mathcal{O}(1/L)$, U-Avg or W-Avg should be preferred over L-Avg, while the latter is preferable for small step sizes $\mathcal{O}(1/L\kappa)$.

4.2 Adjusting m with BB Step Sizes

Since the BB step size can change over a wide range of values (cf. (8)), it is hard to find a single m suitable for both small and large η^s at the same time. From a theoretical perspective, choosing $m = \mathcal{O}(\kappa^2)$ in both Propositions 1 and 2 is mainly for coping with the small step sizes $\eta^s = \mathcal{O}(1/(L\theta_\kappa))$. But such a choice is too pessimistic for large step sizes $\eta^s = \mathcal{O}(1/(\mu\theta_\kappa))$. In fact, choosing $m = \mathcal{O}(\kappa)$ for $\eta^s = \mathcal{O}(1/L)$ is good enough, as suggested by Corollaries 1 and 2. In this subsection, our goal is to design an m^s that changes dynamically per outer loop s.

Reflecting on the convergence of SVRG and SARAH, it is sufficient to set the inner loop length m^s according to the η^s used. To highlight the rationale behind our choice of m^s, let us consider BB-SARAH with U-Avg as an example that features convergence rate $\lambda^s = \frac{1}{\mu\eta^s m^s} + \frac{\eta^s L}{2-\eta^s L}$ [Nguyen et al., 2017]. Set $\theta_\kappa > \kappa$ as in Proposition 2 so that the second term of λ^s is always less than 1. With a large step size $\eta^s = \mathcal{O}(1/L)$, and by simply choosing $m^s = \mathcal{O}\big(1/(\mu\eta^s)\big)$, one can ensure a convergent iteration having e.g., $\lambda^s < 1$. With a small step size $\eta^s = \mathcal{O}(1/\kappa L)$ though, choosing $m^s = \mathcal{O}\big(1/(\mu\eta^s)\big)$ also leads to $\lambda^s < 1$. These considerations prompted us to adopt the η^s in (7), and per outer loop s the number of inner loop iterations as

$$m^s = \frac{c}{\mu\eta^s} \,. \tag{9}$$

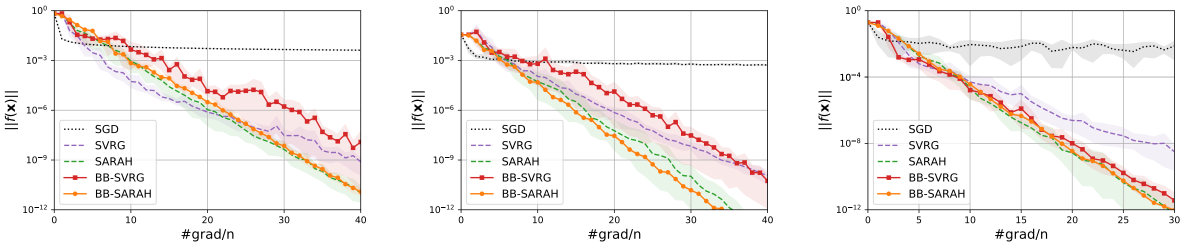
Such choices of η^s and m^s at first glance do not lead to a tune-free algorithm directly, because one has to find an optimal θ_κ and c through tuning. Fortunately, there are simple choices for both c and θ_κ. In Propositions 1 and 2, the smallest selected θ_κ for SVRG and SARAH with different types of averaging turns out to be a reliable choice; while choosing $c = 1$ has been good enough throughout our numerical experiments. Although the selection of these parameters violates slightly the theoretical guarantee, it alleviates the requirement for tuning step sizes and m. And in our experiments, no divergence has been observed by these parameter selections.

5 Numerical Tests

To assess performance, the proposed tune-free BB-SVRG and BB-SARAH are applied to binary classification tasks (cf. (6)) using the datasets *a9a*, *rcv1.binary*, and *real-sim*[3]. Details regarding the datasets, the μ values used, as well as implementation details are deferred to Appendix D.2.

For comparison, the selected benchmarks are SGD, SVRG with U-Avg, and SARAH with U-Avg. The step size for SGD is $\eta = 0.1/(L(n_e + 1))$, where n_e is the index of epochs. For SVRG and SARAH, we fix $m = 5\kappa$, and tune for the best step sizes. For BB-SVRG, we choose η^s and m^s as (9) with $\theta_\kappa = 4\kappa$ (as in Proposition 1) and $c = 1$. While we choose $\theta_\kappa = 3\kappa/2$ (as in Proposition 2) and $c = 1$ for BB-SARAH. As for averaging techniques, W-Avg is adopted when $\eta^s > 0.01/L$, otherwise L-Avg is adopted.

[3] All these datasets are from LIBSVM that is online available at `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`.



(a) *a9a* (b) *rcv1* (c) *real-sim*

Figure 5: Tests of BB-SVRG and BB-SARAH

The results are showcased in Fig. 5. On dataset *a9a*, tuned SARAH and SVRG exhibit improved performance over their BB counterparts. And the gap between BB-SARAH and SARAH tends to be smaller than that between BB-SVRG and SVRG. The price paid for tune-free variance reduction is slightly worse convergence. On dataset *rcv1* however, BB-SARAH outperforms SARAH because $m = 5\kappa$ is not the best choice for m, which suggests that an automatically tuned m^s in (9) can sometimes improve the performance of a suboptimum m in SARAH. BB-SVRG is worse than SVRG initially, but has similar performance around the 40th datum on the x-axis. On dataset *real-sim*, BB-SARAH performs almost identical to SARAH. BB-SVRG exhibits comparable performance with BB-SARAH, but outperforms SVRG.

6 Conclusions

Almost tune-free SVRG and SARAH algorithms were developed in this work. The BB step size is at the core of such tune-free variance reduction algorithms. The key insights are that both i) averaging, as well as ii) the number of inner loop iterations should be adjusted according to the BB step size. Specific major findings include: i) estimate sequence based provably linear convergence of SVRG and SARAH, which enabled novel types of averaging for efficient variance reduction; ii) theoretical guarantees of BB-SVRG and BB-SARAH with different types of averaging; and, iii) implementable tune-free variance reduction algorithms. The efficacy of the novel tune-free BB-SVRG and BB-SARAH were corroborated numerically on real datasets.

References

Alekh Agarwal and Leon Bottou. A lower bound for the optimization of finite sums. In *Proc. Intl. Conf. on Machine Learning*, pages 78–86, Lille, France, 2015.

Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *Proc. Intl. Conf. on Machine Learning*, pages 699–707, New York City, NY, 2016.

Jonathan Barzilai and Jonathan M Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*, 2016.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Proc. Advances in Neural Info. Process. Syst.*, pages 1646–1654, Montreal, Canada, 2014.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Proc. Advances in Neural Info. Process. Syst.*, pages 687–697, Montreal, Canada, 2018.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Proc. Advances in Neural Info. Process. Syst.*, pages 315–323, Lake Tahoe, Nevada, 2013.

Jakub Konecnỳ and Peter Richtárik. Semi-stochastic gradient descent methods. *arXiv preprint arXiv:1312.1666*, 2013.

Dmitry Kovalev, Samuel Horvath, and Peter Richtarik. Don't jump through hoops and remove those loops: Svrg and katyusha are better without the outer loop. *arXiv preprint arXiv:1901.08689*, 2019.

Andrei Kulunchakov and Julien Mairal. Estimate sequences for variance-reduced stochastic composite optimization. *arXiv preprint arXiv:1905.02374*, 2019.

Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via SCSG methods. In *Proc. Advances in Neural Info. Process. Syst.*, pages 2348–2358, 2017.

Bingcong Li, Meng Ma, and Georgios B Giannakis. On the convergence of SARAH and beyond. *arXiv preprint arXiv:1906.02351*, 2019.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Proc. Advances in Neural Info. Process. Syst.*, pages 3384–3392, Montreal, Canada, 2015.

Yan Liu, Congying Han, and Tiande Guo. A class of stochastic variance reduced methods with an adaptive stepsize. URL `http://www.optimization-online.org/DB_FILE/2019/04/7170.pdf`.

Julien Mairal. Optimization with first-order surrogate functions. In *Proc. Intl. Conf. on Machine Learning*, pages 783–791, Atlanta, 2013.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2004.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proc. Intl. Conf. Machine Learning*, Sydney, Australia, 2017.

Lam M Nguyen, Marten van Dijk, Dzung T Phan, Phuong Ha Nguyen, Tsui-Wei Weng, and Jayant R Kalagnanam. Optimal finite-sum smooth non-convex optimization with SARAH. *arXiv preprint arXiv:1901.07648*, 2019.

Atsushi Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *Proc. Advances in Neural Info. Process. Syst.*, pages 1574–1582, Montreal, Canada, 2014.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *Proc. Intl. Conf. on Machine Learning*, pages 314–323, New York City, NY, 2016.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Nicolas L Roux, Mark Schmidt, and Francis R Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Proc. Advances in Neural Info. Process. Syst.*, pages 2663–2671, Lake Tahoe, Nevada, 2012.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

Conghui Tan, Shiqian Ma, Yu-Hong Dai, and Yuqiu Qian. Barzilai-Borwein step size for stochastic gradient descent. In *Proc. Advances in Neural Info. Process. Syst.*, pages 685–693, 2016.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Zhuang Yang, Zengping Chen, and Cheng Wang. Accelerating mini-batch sarah by step size rules. *arXiv preprint arXiv:1906.08496*, 2019.

Supplementary Document for "Almost Tune-Free Variance Reduction"

A Properties of ES

A.1 Proof of Lemma 2

i) By definition $\Phi_0(\mathbf{x})$ is μ_0-strongly convex; and $\Phi_k(\mathbf{x})$ is μ_k-strongly convex with $\mu_k = (1-\delta_k)\mu_{k-1} + \delta_k \mu$.

ii) Clearly, $\mathbf{x}_0$ minimizes $\Phi_0(\mathbf{x})$. Arguing by induction, suppose that $\mathbf{x}_{k-1}$ minimizes $\Phi_{k-1}(\mathbf{x})$, to obtain

$$\Phi_{k-1}(\mathbf{x}) = \Phi_{k-1}^* + \frac{\mu_{k-1}}{2}\|\mathbf{x} - \mathbf{x}_{k-1}\|^2 \quad \Rightarrow \quad \nabla\Phi_{k-1}(\mathbf{x}) = \mu_{k-1}(\mathbf{x} - \mathbf{x}_{k-1}).$$

By definition of $\Phi_k(\mathbf{x})$, we also have

$$\begin{aligned}\nabla\Phi_k(\mathbf{x}) &= (1-\delta_k)\nabla\Phi_{k-1}(\mathbf{x}) + \delta_k \mathbf{v}_{k-1} + \mu\delta_k(\mathbf{x} - \mathbf{x}_{k-1}) \\ &= (1-\delta_k)\mu_{k-1}(\mathbf{x} - \mathbf{x}_{k-1}) + \delta_k \mathbf{v}_{k-1} + \mu\delta_k(\mathbf{x} - \mathbf{x}_{k-1}). \end{aligned} \tag{10}$$

Using $\mu_k = (1-\delta_k)\mu_{k-1} + \delta_k\mu$ and setting $\nabla\Phi_k(\mathbf{x}) = \mathbf{0}$, we find that $\mathbf{x}_k$ minimizes $\Phi_k(\mathbf{x})$ when $\delta_k = \eta\mu_k$.

iii) Since $\mathbf{x}_{k-1}$ minimizes $\Phi_{k-1}(\mathbf{x})$, using the definition of $\Phi_k(\mathbf{x})$ we can write

$$\Phi_k(\mathbf{x}_{k-1}) = (1-\delta_k)\Phi_{k-1}^* + \delta_k f(\mathbf{x}_{k-1}). \tag{11}$$

On the other hand, we also have $\Phi_k(\mathbf{x}_{k-1}) = \Phi_k^* + \frac{\mu_k}{2}\|\mathbf{x}_{k-1} - \mathbf{x}_k\|^2$. Comparing this with (11) and using that $\mathbf{x}_k = \mathbf{x}_{k-1} - \eta\mathbf{v}_{k-1}$, completes the proof of this property.

A.2 Derivations of (4) and (5)

To verify (4), proceed as follows

$$\begin{aligned}\mathbb{E}\big[\Phi_k(\mathbf{x})\big] &= (1-\delta)\mathbb{E}\big[\Phi_{k-1}(\mathbf{x})\big] + \delta\mathbb{E}\bigg[f(\mathbf{x}_{k-1}) + \langle \mathbf{v}_{k-1}, \mathbf{x} - \mathbf{x}_{k-1}\rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}_{k-1}\|^2\bigg] \\ &= (1-\delta)\mathbb{E}\big[\Phi_{k-1}(\mathbf{x})\big] + \delta\mathbb{E}\bigg[f(\mathbf{x}_{k-1}) + \langle \nabla f(\mathbf{x}_{k-1}), \mathbf{x} - \mathbf{x}_{k-1}\rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}_{k-1}\|^2\bigg] \\ &\leq (1-\delta)\mathbb{E}\big[\Phi_{k-1}(\mathbf{x})\big] + \delta f(\mathbf{x}) \leq (1-\delta)^k\big[\Phi_0(\mathbf{x}) - f(\mathbf{x})\big] + f(\mathbf{x}) \leq (1-\delta)^k\big[\Phi_0(\mathbf{x}) - f(\mathbf{x}^*)\big] + f(\mathbf{x}). \end{aligned} \tag{12}$$

And in order to derive (5), follow the next steps

$$\begin{aligned}\mathbb{E}\big[\Phi_k(\mathbf{x})\big] &= (1-\delta)\mathbb{E}\big[\Phi_{k-1}(\mathbf{x})\big] + \delta\mathbb{E}\bigg[f(\mathbf{x}_{k-1}) + \langle \mathbf{v}_{k-1}, \mathbf{x} - \mathbf{x}_{k-1}\rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}_{k-1}\|^2\bigg] \\ &\leq (1-\delta)\mathbb{E}\big[\Phi_{k-1}(\mathbf{x})\big] + \delta f(\mathbf{x}) + \delta\mathbb{E}\big[\langle \mathbf{v}_{k-1} - \nabla f(\mathbf{x}_{k-1}), \mathbf{x} - \mathbf{x}_{k-1}\rangle\big] \\ &\leq (1-\delta)^k\big[\Phi_0(\mathbf{x}) - f(\mathbf{x})\big] + f(\mathbf{x}) + \underbrace{\delta\sum_{\tau=0}^{k-1}(1-\delta)^\tau\mathbb{E}\big[\langle \mathbf{v}_{k-1-\tau} - \nabla f(\mathbf{x}_{k-1-\tau}), \mathbf{x} - \mathbf{x}_{k-1-\tau}\rangle\big]}_{:=C;\ C\neq 0,\ \text{extra term compared with SVRG}}. \end{aligned}$$

A.3 A key lemma

The next lemma plays a major role in our analysis.

Lemma 4. *If we choose $\mu_0 = \mu$, $\delta_k = \mu_k\eta$, and $\Phi_0^* = f(\mathbf{x}_0)$ in the ES defined in* (3)*, we then find that: i) $\mu_k = \mu, \forall k$; ii) $\delta := \delta_k = \mu\eta$; and iii) the following inequality holds*

$$\begin{aligned}&\delta\sum_{\tau=1}^{k-1}(1-\delta)^{k-\tau-1}\big[f(\mathbf{x}_\tau) - f(\mathbf{x}^*)\big] + (1-\delta)^{k-1}\big[f(\mathbf{x}_0) - f(\mathbf{x}^*)\big] \\ &\leq (1-\delta)^k\big[\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)\big] + \frac{\mu\eta^2}{2}\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\|\mathbf{v}_{\tau-1}\|^2 + \delta\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\zeta_{\tau-1}\end{aligned}$$

where $\zeta_{k-1} := \langle \mathbf{v}_{k-1} - \nabla f(\mathbf{x}_{k-1}), \mathbf{x}^* - \mathbf{x}_{k-1}\rangle$.

Proof. Since i) and ii) are straightforward to verify, we will prove iii). Using property iii) in Lemma 2, we find

$$\begin{aligned}
f(\mathbf{x}_k) - \Phi_k^* &= f(\mathbf{x}_k) - (1-\delta_k)\Phi_{k-1}^* - \delta_k f(\mathbf{x}_{k-1}) + \frac{\mu_k \eta^2}{2}\|\mathbf{v}_{k-1}\|^2 \\
&= f(\mathbf{x}_k) - \Phi_{k-1}^* + \delta_k\big(\Phi_{k-1}^* - f(\mathbf{x}_{k-1})\big) + \frac{\mu_k \eta^2}{2}\|\mathbf{v}_{k-1}\|^2 \\
&= f(\mathbf{x}_k) - f(\mathbf{x}_{k-1}) + f(\mathbf{x}_{k-1}) - \Phi_{k-1}^* + \delta_k\big(\Phi_{k-1}^* - f(\mathbf{x}_{k-1})\big) + \frac{\mu_k \eta^2}{2}\|\mathbf{v}_{k-1}\|^2 \\
&= (1-\delta_k)\big[f(\mathbf{x}_{k-1}) - \Phi_{k-1}^*\big] + \xi_k \qquad (13)
\end{aligned}$$

where ξ_k is defined as

$$\xi_k := f(\mathbf{x}_k) - f(\mathbf{x}_{k-1}) + \frac{\mu_k \eta^2}{2}\|\mathbf{v}_{k-1}\|^2.$$

Upon expanding $f(\mathbf{x}_{k-1}) - \Phi_{k-1}^*$ in (13), we have

$$\begin{aligned}
f(\mathbf{x}_k) - \Phi_k^* &= (1-\delta_k)\big[f(\mathbf{x}_{k-1}) - \Phi_{k-1}^*\big] + \xi_k \\
&= \Big[\prod_{\tau=1}^{k}(1-\delta_\tau)\Big][f(\mathbf{x}_0) - \Phi_0^*] + \sum_{\tau=1}^{k}\xi_\tau\Big[\prod_{j=\tau+1}^{k}(1-\delta_j)\Big] \qquad (14)
\end{aligned}$$

from which we deduce that

$$\begin{aligned}
\Phi_k^* \le \Phi_k(\mathbf{x}^*) &= (1-\delta_k)\Phi_{k-1}(\mathbf{x}^*) + \delta_k\Big[f(\mathbf{x}_{k-1}) + \langle \mathbf{v}_{k-1}, \mathbf{x}^* - \mathbf{x}_{k-1}\rangle + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_{k-1}\|^2\Big] \\
&\overset{(a)}{=} (1-\delta_k)\Phi_{k-1}(\mathbf{x}^*) + \delta_k\Big[f(\mathbf{x}_{k-1}) + \langle \nabla f(\mathbf{x}_{k-1}), \mathbf{x}^* - \mathbf{x}_{k-1}\rangle + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_{k-1}\|^2 + \zeta_{k-1}\Big] \\
&\overset{(b)}{\le} (1-\delta_k)\Phi_{k-1}(\mathbf{x}^*) + \delta_k f(\mathbf{x}^*) + \delta_k\zeta_{k-1} \\
&\le \Big[\prod_{\tau=1}^{k}(1-\delta_\tau)\Big]\Phi_0(\mathbf{x}^*) + \sum_{\tau=1}^{k}\delta_\tau f(\mathbf{x}^*)\Big[\prod_{j=\tau+1}^{k}(1-\delta_j)\Big] + \sum_{\tau=1}^{k}\delta_\tau\zeta_{\tau-1}\Big[\prod_{j=\tau+1}^{k}(1-\delta_j)\Big] \qquad (15)
\end{aligned}$$

where in (a) the ζ_{k-1} is defined as

$$\zeta_{k-1} := \langle \mathbf{v}_{k-1} - \nabla f(\mathbf{x}_{k-1}), \mathbf{x}^* - \mathbf{x}_{k-1}\rangle;$$

and (b) follows from the strongly convexity of f. Then, using (14), we have

$$\begin{aligned}
f(\mathbf{x}_k) - f(\mathbf{x}^*) &= \Phi_k^* - f(\mathbf{x}^*) + \Big[\prod_{\tau=1}^{k}(1-\delta_\tau)\Big][f(\mathbf{x}_0) - \Phi_0^*] + \sum_{\tau=1}^{k}\xi_\tau\Big[\prod_{j=\tau+1}^{k}(1-\delta_j)\Big] \\
&\overset{(c)}{\le} \Big[\prod_{\tau=1}^{k}(1-\delta_\tau)\Big]\Phi_0(\mathbf{x}^*) + \sum_{\tau=1}^{k}\delta_\tau f(\mathbf{x}^*)\Big[\prod_{j=\tau+1}^{k}(1-\delta_j)\Big] + \sum_{\tau=1}^{k}\delta_\tau\zeta_{\tau-1}\Big[\prod_{j=\tau+1}^{k}(1-\delta_j)\Big] \\
&\qquad - f(\mathbf{x}^*) + \Big[\prod_{\tau=1}^{k}(1-\delta_\tau)\Big][f(\mathbf{x}_0) - \Phi_0^*] + \sum_{\tau=1}^{k}\xi_\tau\Big[\prod_{j=\tau+1}^{k}(1-\delta_j)\Big]
\end{aligned}$$

where (c) is due to (15). Choosing $\mu_0 = \mu$ (hence $\mu_k = \mu, \delta_k = \mu\eta := \delta,\ \forall k$) and $\Phi_0^* = f(\mathbf{x}_0)$, we arrive at

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \le (1-\delta)^k\big[\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)\big] + \sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\big(\xi_\tau + \delta\zeta_{\tau-1}\big). \qquad (16)$$

Now consider that

$$\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\xi_\tau = \sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\Big[f(\mathbf{x}_\tau) - f(\mathbf{x}_{\tau-1}) + \frac{\mu\eta^2}{2}\|\mathbf{v}_{\tau-1}\|^2\Big]$$
$$= f(\mathbf{x}_k) + \sum_{\tau=1}^{k-1}(1-\delta)^{k-\tau}f(\mathbf{x}_\tau) - \sum_{\tau=1}^{k-1}(1-\delta)^{k-\tau-1}f(\mathbf{x}_\tau) - (1-\delta)^{k-1}f(\mathbf{x}_0) + \frac{\mu\eta^2}{2}\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\|\mathbf{v}_{\tau-1}\|^2$$
$$= -\delta\sum_{\tau=1}^{k-1}(1-\delta)^{k-\tau-1}f(\mathbf{x}_\tau) + f(\mathbf{x}_k) - (1-\delta)^{k-1}f(\mathbf{x}_0) + \frac{\mu\eta^2}{2}\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\|\mathbf{v}_{\tau-1}\|^2. \tag{17}$$

Because $\delta\sum_{\tau=1}^{k-1}(1-\delta)^{k-\tau-1} + (1-\delta)^{k-1} = 1$, we can write $f(\mathbf{x}^*) = [\delta\sum_{\tau=1}^{k-1}(1-\delta)^{k-\tau-1} + (1-\delta)^{k-1}]f(\mathbf{x}^*)$. Using the latter, plugging (17) into (16), and eliminating $f(\mathbf{x}_k)$, we obtain

$$\delta\sum_{\tau=1}^{k-1}(1-\delta)^{k-\tau-1}\big[f(\mathbf{x}_\tau) - f(\mathbf{x}^*)\big] + (1-\delta)^{k-1}\big[f(\mathbf{x}_0) - f(\mathbf{x}^*)\big]$$
$$\leq (1-\delta)^k\big[\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)\big] + \frac{\mu\eta^2}{2}\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\|\mathbf{v}_{\tau-1}\|^2 + \delta\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\zeta_{\tau-1} \tag{18}$$

which completes the proof. □

B Proofs for SVRG and SARAH

B.1 Proof for SVRG (Theorem 1 and Corollary 1)

Proof of Theorem 1

Proof. Since the choices of μ_0, Φ_0^* and δ_k coincide with those in Lemma 4, we can directly apply Lemma 4 to find

$$\delta\sum_{\tau=1}^{k-1}(1-\delta)^{k-\tau-1}\big[f(\mathbf{x}_\tau) - f(\mathbf{x}^*)\big] + (1-\delta)^{k-1}\big[f(\mathbf{x}_0) - f(\mathbf{x}^*)\big]$$
$$\leq (1-\delta)^k\big[\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)\big] + \frac{\mu\eta^2}{2}\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\|\mathbf{v}_{\tau-1}\|^2 + \delta\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\zeta_{\tau-1} \tag{19}$$

where $\zeta_{k-1} := \langle\mathbf{v}_{k-1} - \nabla f(\mathbf{x}_{k-1}), \mathbf{x}^* - \mathbf{x}_{k-1}\rangle$. Upon defining the σ-algebra $\mathcal{F}_{k-1} = \sigma(i_0, i_1, \dots, i_{k-1})$, and using that $\mathbf{v}_k$ is an unbiased estimator of $\nabla f(\mathbf{x}_k)$, it follows readily that

$$\mathbb{E}[\zeta_k|\mathcal{F}_{k-1}] = \mathbb{E}\big[\mathbf{v}_k - \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k\rangle|\mathcal{F}_{k-1}\big] = 0$$

which further implies

$$\mathbb{E}[\zeta_k] = 0. \tag{20}$$

Now taking expectation on both sides of (19) and using (20), we have

$$\delta\sum_{\tau=1}^{k-1}(1-\delta)^{k-\tau-1}\mathbb{E}\big[f(\mathbf{x}_\tau) - f(\mathbf{x}^*)\big] + (1-\delta)^{k-1}\mathbb{E}\big[f(\mathbf{x}_0) - f(\mathbf{x}^*)\big] \tag{21}$$
$$\leq (1-\delta)^k\mathbb{E}\big[\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)\big] + \frac{\mu\eta^2}{2}\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\mathbb{E}\big[\|\mathbf{v}_{\tau-1}\|^2\big]$$
$$\overset{(a)}{\leq} (1-\delta)^k\mathbb{E}\big[\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)\big] + 2\mu L\eta^2\sum_{\tau=0}^{k-1}(1-\delta)^{k-\tau-1}\mathbb{E}\big[f(\mathbf{x}_\tau) - f(\mathbf{x}^*) + f(\mathbf{x}_0) - f(\mathbf{x}^*)\big]$$
$$\overset{(b)}{\leq} (1-\delta)^k\mathbb{E}\big[\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)\big] + 2\mu L\eta^2\sum_{\tau=0}^{k-1}(1-\delta)^{k-\tau-1}\mathbb{E}\big[f(\mathbf{x}_\tau) - f(\mathbf{x}^*)\big] + \frac{2\mu L\eta^2}{\delta}\mathbb{E}\big[f(\mathbf{x}_0) - f(\mathbf{x}^*)\big]$$

where in (a) we used Lemma 1 to $\mathbb{E}[\|\mathbf{v}_{\tau-1}\|^2]$; and (b) holds because $\sum_{\tau=0}^{k-1}(1-\delta)^{k-\tau-1} \leq 1/\delta$. Note that we can use $\Phi_0(\mathbf{x}^*) = f(\mathbf{x}_0) + \frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|^2$ together with $(1-\delta)^{k-1} > (1-\delta)^k$, to eliminate $(1-\delta)^{k-1}\mathbb{E}[f(\mathbf{x}_0) - f(\mathbf{x}^*)]$ on the LHS of (21). Rearranging the terms, we arrive at

$$
\begin{aligned}
&(\delta - 2\mu L\eta^2)\sum_{\tau=1}^{k-1}(1-\delta)^{k-\tau-1}\mathbb{E}\big[f(\mathbf{x}_\tau) - f(\mathbf{x}^*)\big] \\
&\leq \frac{\mu}{2}(1-\delta)^k\mathbb{E}\big[\|\mathbf{x}_0 - \mathbf{x}^*\|^2\big] + 2\mu L\eta^2(1-\delta)^{k-1}\mathbb{E}\big[f(\mathbf{x}_0) - f(\mathbf{x}^*)\big] + \frac{2\mu L\eta^2}{\delta}\mathbb{E}\big[f(\mathbf{x}_0) - f(\mathbf{x}^*)\big] \\
&\leq \bigg[(1-\delta)^k + 2\mu L\eta^2(1-\delta)^{k-1} + \frac{2\mu L\eta^2}{\delta}\bigg]\mathbb{E}\big[f(\mathbf{x}_0) - f(\mathbf{x}^*)\big] \qquad (22)
\end{aligned}
$$

where the last inequality is due to $\frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^*\| \leq f(\mathbf{x}) - f(\mathbf{x}^*)$. Now choosing $\eta < 1/2L$ so that $\delta - 2\mu L\eta^2 > 0$, we have

$$
\sum_{\tau=1}^{k-1}(1-\delta)^{k-\tau-1}\mathbb{E}\big[f(\mathbf{x}_\tau) - f(\mathbf{x}^*)\big] \leq \bigg[\frac{(1-\delta)^k}{\delta - 2\mu L\eta^2} + \frac{2\mu L\eta^2(1-\delta)^{k-1}}{\delta - 2\mu L\eta^2} + \frac{2\mu L\eta^2}{\delta(\delta - 2\mu L\eta^2)}\bigg]\mathbb{E}\big[f(\mathbf{x}_0) - f(\mathbf{x}^*)\big].
$$

With $p_0 = p_m = 0$, and $p_k = (1-\delta)^{m-k-1}/q, k = 1, 2, \ldots, m-1$, where $q = [1-(1-\delta)^{m-1}]/\delta$ (with $\delta = \mu\eta$), we find

$$
\begin{aligned}
&\mathbb{E}\big[f(\tilde{\mathbf{x}}^s) - f(\mathbf{x}^*)\big] = \sum_{\tau=1}^{m-1}\frac{(1-\delta)^{m-\tau-1}}{q}\mathbb{E}\big[f(\mathbf{x}_\tau) - f(\mathbf{x}^*)\big] \\
&\leq \frac{1}{q}\bigg[\frac{(1-\delta)^m}{\delta - 2\mu L\eta^2} + \frac{2\mu L\eta^2(1-\delta)^{m-1}}{\delta - 2\mu L\eta^2} + \frac{2\mu L\eta^2}{\delta(\delta - 2\mu L\eta^2)}\bigg]\mathbb{E}\big[f(\tilde{\mathbf{x}}^{s-1}) - f(\mathbf{x}^*)\big] \\
&= \underbrace{\frac{1}{1-(1-\mu\eta)^{m-1}}\bigg[\frac{(1-\mu\eta)^m}{1-2\eta L} + \frac{2\mu L\eta^2(1-\mu\eta)^{m-1}}{1-2L\eta} + \frac{2L\eta}{1-2L\eta}\bigg]}_{:=\lambda^{\text{SVRG}}}\mathbb{E}\big[f(\tilde{\mathbf{x}}^{s-1}) - f(\mathbf{x}^*)\big]. \qquad (23)
\end{aligned}
$$

Thus, so long as we choose a large enough m and $\eta < 1/(4L)$, we have $\lambda^{\text{SVRG}} < 1$, that is, SVRG converges linearly. □

Proof of Corollary 1

Proof. Choose $\eta = 1/(8L)$ and $m = \frac{3}{\mu\eta} + 1 = 24\kappa + 1 \geq 25$. We have that

$$
(1-\mu\eta)^{\frac{1}{\mu\eta}} \leq 0.4 \quad \Rightarrow \quad (1-\mu\eta)^m \leq (0.4)^3
$$

(Actually $(1-\mu\eta)^{\frac{1}{\mu\eta}} \approx 1/e$ when $\mu\eta$ small enough). Using the value of η and m, it can be verified that $\lambda^{\text{SVRG}} \leq 0.5$. This implies that $\mathcal{O}\big(\ln\frac{1}{\epsilon}\big)$ outer loops are needed for an ϵ-accurate solution. And since $m = \mathcal{O}(\kappa)$, the overall IFO complexity is $\mathcal{O}\big((n+\kappa)\ln\frac{1}{\epsilon}\big)$. □

B.2 Proofs for SARAH (Lemma 3, Theorem 2 and Corollary 2)

Proof of Lemma 3

Proof. Let $\mathcal{F}_{k-1} = \sigma(i_1, i_2, \ldots, i_{k-1})$, then for any $\mathbf{x}$ we have

$$
\begin{aligned}
\mathbb{E}\big[\langle\mathbf{v}_k - f(\mathbf{x}_k),\, \mathbf{x} - \mathbf{x}_k\rangle|\mathcal{F}_{k-1}\big] &= \mathbb{E}\big[\langle\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}_{k-1}) + \mathbf{v}_{k-1} - \nabla f(\mathbf{x}_k),\, \mathbf{x} - \mathbf{x}_k\rangle|\mathcal{F}_{k-1}\big] \\
&= \langle\mathbf{v}_{k-1} - \nabla f(\mathbf{x}_{k-1}),\, \mathbf{x} - \mathbf{x}_k\rangle = \langle\mathbf{v}_{k-1} - \nabla f(\mathbf{x}_{k-1}),\, \mathbf{x} - \mathbf{x}_{k-1} + \mathbf{x}_{k-1} - \mathbf{x}_k\rangle \\
&= \langle\mathbf{v}_{k-1} - \nabla f(\mathbf{x}_{k-1}),\, \mathbf{x} - \mathbf{x}_{k-1}\rangle + \eta\langle\mathbf{v}_{k-1} - \nabla f(\mathbf{x}_{k-1}),\, \mathbf{v}_{k-1}\rangle \\
&= \langle\mathbf{v}_{k-1} - \nabla f(\mathbf{x}_{k-1}),\, \mathbf{x} - \mathbf{x}_{k-1}\rangle + \frac{\eta}{2}\Big[\|\mathbf{v}_{k-1}\|^2 + \|\mathbf{v}_{k-1} - \nabla f(\mathbf{x}_{k-1})\|^2 - \|\nabla f(\mathbf{x}_{k-1})\|^2\Big]
\end{aligned}
$$

where the last equation is because $2\langle\mathbf{a}, \mathbf{b}\rangle = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2$. Since $\mathbf{v}_0 = \nabla f(\mathbf{x}_0)$, we have $\langle\mathbf{v}_0 - \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0\rangle = 0$. And the proof is completed, after taking expectation and unrolling $\langle\mathbf{v}_{k-1} - \nabla f(\mathbf{x}_{k-1}), \mathbf{x} - \mathbf{x}_{k-1}\rangle$. □

In order to prove Theorem 2, we need to borrow the following result from [Nguyen et al., 2017].

Lemma 5. *[Nguyen et al., 2017, Theorem 1b] If Assumptions 1 and 4 hold, with $\eta \leq 2/(\mu+L)$, SARAH guarantees*

$$\mathbb{E}\big[\|\mathbf{v}_k\|^2\big] \leq \left(1 - \frac{2\eta L}{1+\kappa}\right)^k \mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big].$$

Proof of Theorem 2.

Proof. With the choices of μ_0, Φ_0^* and δ_k as in Lemma 4, we can directly apply Lemma 4 to confirm that

$$\begin{aligned}
&(1-\delta)^{k-1}\big[f(\mathbf{x}_0) - f(\mathbf{x}^*)\big] + \delta \sum_{\tau=1}^{k-1}(1-\delta)^{k-\tau-1}\big[f(\mathbf{x}_\tau) - f(\mathbf{x}^*)\big] \\
&\leq (1-\delta)^k\big[\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)\big] + \frac{\mu\eta^2}{2}\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\|\mathbf{v}_{\tau-1}\|^2 + \sum_{\tau=1}^{k}\delta(1-\delta)^{k-\tau}\langle \mathbf{v}_{\tau-1} - \nabla f(\mathbf{x}_{\tau-1}), \mathbf{x}^* - \mathbf{x}_{\tau-1}\rangle \\
&= (1-\delta)^k\big[\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)\big] + \frac{\mu\eta^2}{2}\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\|\mathbf{v}_{\tau-1}\|^2 + \sum_{\tau=2}^{k}\delta(1-\delta)^{k-\tau}\langle \mathbf{v}_{\tau-1} - \nabla f(\mathbf{x}_{\tau-1}), \mathbf{x}^* - \mathbf{x}_{\tau-1}\rangle
\end{aligned}$$

where the last equation holds because $\mathbf{v}_0 = \nabla f(\mathbf{x}_0)$. Since $\Phi_0(\mathbf{x}^*) = f(\mathbf{x}_0) + \frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|^2 \leq f(\mathbf{x}_0) + \frac{1}{2\mu}\|\nabla f(\mathbf{x}_0)\|^2$ and $(1-\delta)^{k-1} > (1-\delta)^k$, we can eliminate $(1-\delta)^{k-1}\mathbb{E}[f(\mathbf{x}_0) - f(\mathbf{x}^*)]$ on the LHS, to obtain the inequality

$$\begin{aligned}
&\delta\sum_{\tau=1}^{k-1}(1-\delta)^{k-\tau-1}\big[f(\mathbf{x}_\tau) - f(\mathbf{x}^*)\big] \\
&\leq \frac{(1-\delta)^k}{2\mu}\|\nabla f(\mathbf{x}_0)\|^2 + \frac{\mu\eta^2}{2}\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\|\mathbf{v}_{\tau-1}\|^2 + \sum_{\tau=2}^{k}\delta(1-\delta)^{k-\tau}\langle \mathbf{v}_{\tau-1} - \nabla f(\mathbf{x}_{\tau-1}), \mathbf{x}^* - \mathbf{x}_{\tau-1}\rangle.
\end{aligned}$$

Taking expectation on both sides, we arrive at

$$\begin{aligned}
&0 \leq \delta\sum_{\tau=1}^{k-1}(1-\delta)^{k-\tau-1}\mathbb{E}\big[f(\mathbf{x}_\tau) - f(\mathbf{x}^*)\big] &(24)\\
&\leq \frac{(1-\delta)^k}{2\mu}\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big] + \frac{\mu\eta^2}{2}\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\mathbb{E}\big[\|\mathbf{v}_{\tau-1}\|^2\big] + \sum_{\tau=2}^{k}\delta(1-\delta)^{k-\tau}\mathbb{E}\big[\langle \mathbf{v}_{\tau-1} - \nabla f(\mathbf{x}_{\tau-1}), \mathbf{x}^* - \mathbf{x}_{\tau-1}\rangle\big] \\
&= \frac{(1-\delta)^k}{2\mu}\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big] + \frac{\mu\eta^2}{2}\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\mathbb{E}\big[\|\mathbf{v}_{\tau-1}\|^2\big] + \sum_{\tau=1}^{k-1}\delta(1-\delta)^{k-1-\tau}\mathbb{E}\big[\langle \mathbf{v}_\tau - \nabla f(\mathbf{x}_\tau), \mathbf{x}^* - \mathbf{x}_\tau\rangle\big] \\
&\leq \frac{(1-\delta)^k}{2\mu}\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big] + \frac{\mu\eta^2}{2}\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\mathbb{E}\big[\|\mathbf{v}_{\tau-1}\|^2\big] \\
&\qquad + \frac{\delta\eta}{2}\sum_{\tau=1}^{k-1}(1-\delta)^{k-1-\tau}\sum_{j=0}^{\tau-1}\mathbb{E}\Big[\|\mathbf{v}_j - \nabla f(\mathbf{x}_j)\|^2 + \|\mathbf{v}_j\|^2 - \|\nabla f(\mathbf{x}_j)\|^2\Big]
\end{aligned}$$

where for the last inequality we used Lemma 3. Changing the summation order in the last term of the RHS of (24), yields

$$\begin{aligned}
&\frac{\delta\eta}{2}\sum_{\tau=1}^{k-1}(1-\delta)^{k-1-\tau}\sum_{j=0}^{\tau-1}\mathbb{E}\Big[\|\mathbf{v}_j - \nabla f(\mathbf{x}_j)\|^2 + \|\mathbf{v}_j\|^2 - \|\nabla f(\mathbf{x}_j)\|^2\Big] \\
&= \frac{\delta\eta}{2}\sum_{\tau=0}^{k-2}\mathbb{E}\Big[\|\mathbf{v}_\tau - \nabla f(\mathbf{x}_\tau)\|^2 + \|\mathbf{v}_\tau\|^2 - \|\nabla f(\mathbf{x}_\tau)\|^2\Big]\left[\sum_{j=0}^{k-\tau-2}(1-\delta)^\tau\right] \\
&\leq \frac{\eta}{2}\sum_{\tau=0}^{k-2}\Big(\mathbb{E}\big[\|\mathbf{v}_\tau - \nabla f(\mathbf{x}_\tau)\|^2\big] + \mathbb{E}\big[\|\mathbf{v}_\tau\|^2\big]\Big) - \frac{\eta}{2}\sum_{\tau=0}^{k-2}\big(1-(1-\delta)^{k-\tau-1}\big)\mathbb{E}\big[\|\nabla f(\mathbf{x}_\tau)\|^2\big]. &(25)
\end{aligned}$$

Now plugging (25) into (24), and rearranging the terms, we find

$$\begin{aligned}
&\frac{\eta}{2}\sum_{\tau=0}^{k-2}\big(1-(1-\delta)^{k-1-\tau}\big)\mathbb{E}\big[\|\nabla f(\mathbf{x}_\tau)\|^2\big] \\
&\le \frac{(1-\delta)^k}{2\mu}\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big] + \frac{\mu\eta^2}{2}\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\mathbb{E}\big[\|\mathbf{v}_{\tau-1}\|^2\big] + \frac{\eta}{2}\sum_{\tau=0}^{k-2}\bigg(\mathbb{E}\big[\|\mathbf{v}_\tau-\nabla f(\mathbf{x}_\tau)\|^2\big]+\mathbb{E}\big[\|\mathbf{v}_\tau\|^2\big]\bigg).
\end{aligned}$$

Dividing both sides by $\eta/2$ (and recalling that $\delta=\mu\eta$), we arrive at

$$\begin{aligned}
&\sum_{\tau=0}^{k-2}\big(1-(1-\delta)^{k-\tau-1}\big)\mathbb{E}\big[\|\nabla f(\mathbf{x}_\tau)\|^2\big] \qquad (26)\\
&\le \frac{(1-\delta)^k}{\mu\eta}\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big] + \delta\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\mathbb{E}\big[\|\mathbf{v}_{\tau-1}\|^2\big] + \sum_{\tau=0}^{k-2}\bigg(\mathbb{E}\big[\|\mathbf{v}_\tau-\nabla f(\mathbf{x}_\tau)\|^2\big]+\mathbb{E}\big[\|\mathbf{v}_\tau\|^2\big]\bigg)\\
&\overset{(a)}{\le} \frac{(1-\delta)^k}{\mu\eta}\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big] + \delta\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\mathbb{E}\big[\|\mathbf{v}_{\tau-1}\|^2\big] + \frac{\eta L(k-1)}{2-\eta L}\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big] + \frac{2-2\eta L}{2-\eta L}\sum_{\tau=0}^{k-2}\mathbb{E}\big[\|\mathbf{v}_\tau\|^2\big]\\
&\overset{(b)}{\le} \frac{(1-\delta)^k}{\mu\eta}\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big] + \delta\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\mathbb{E}\big[\|\mathbf{v}_{\tau-1}\|^2\big] + \frac{\eta L(k-1)}{2-\eta L}\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big] + \frac{2-2\eta L}{2-\eta L}\frac{1+\kappa}{2\eta L}\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big]\\
&\overset{(c)}{\le} \frac{(1-\delta)^k}{\mu\eta}\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big] + \bigg[(1-\delta)^k-\Big(1-\frac{2\eta L}{1+\kappa}\Big)^k\bigg]\frac{L+\mu}{L-\mu}\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big]\\
&\qquad\qquad + \frac{\eta L(k-1)}{2-\eta L}\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big] + \frac{2-2\eta L}{2-\eta L}\frac{1+\kappa}{2L\eta}\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big]
\end{aligned}$$

where in (a) we applied Lemma 1 to deal with $\mathbb{E}[\|\nabla f(\mathbf{x}_\tau)-\mathbf{v}_\tau\|^2]$; in (b) we chose $\eta<1/L$ and used Lemma 5 to handle $\mathbb{E}[\|\mathbf{v}_\tau\|^2]$ in the last term; and the derivation of (c) is as follows. First, notice that $2\eta L/(1+\kappa)>\mu\eta=\delta$, which implies that $1-\delta>1-[2\eta L/(1+\kappa)]$. Then, leveraging Lemma 5, we have

$$\begin{aligned}
\delta\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\mathbb{E}\big[\|\mathbf{v}_{\tau-1}\|^2\big] &\le \delta\sum_{\tau=1}^{k}(1-\delta)^{k-\tau}\bigg(1-\frac{2\eta L}{1+\kappa}\bigg)^{\tau-1}\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big]\\
&= \bigg[(1-\delta)^k-\Big(1-\frac{2\eta L}{1+\kappa}\Big)^k\bigg]\frac{L+\mu}{L-\mu}\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big].
\end{aligned}$$

To proceed, define

$$c:=\sum_{\tau=0}^{m-2}\big(1-(1-\delta)^{m-\tau-1}\big)=(m-1)-\frac{(1-\delta)-(1-\delta)^m}{\delta}=m-\frac{1}{\delta}+\frac{(1-\delta)^m}{\delta}.$$

and select m large enough so that $c>0$. Upon setting $p_k=(1-(1-\delta)^{m-k-1})/c, \forall k=0,1,\ldots,m-2$, and $p_{m-1}=p_m=0$, we have

$$\begin{aligned}
\mathbb{E}\big[\|\nabla f(\tilde{\mathbf{x}}^s)\|\big] &= \frac{1}{c}\sum_{\tau=0}^{m-2}\big(1-(1-\delta)^{m-\tau-1}\big)\mathbb{E}\big[\|\nabla f(\mathbf{x}_\tau)\|^2\big]\\
&\le \underbrace{\bigg[\frac{(1-\delta)^m}{c\mu\eta}+\bigg((1-\delta)^m-\Big(1-\frac{2\eta L}{1+\kappa}\Big)^m\bigg)\frac{L+\mu}{c(L-\mu)}+\frac{\eta L(m-1)}{c(2-\eta L)}+\frac{2-2\eta L}{2-\eta L}\frac{1+\kappa}{2c\eta L}\bigg]}_{:=\lambda^{\text{SARAH}}}\mathbb{E}\big[\|\nabla f(\tilde{\mathbf{x}}^{s-1})\|^2\big].
\end{aligned}$$

Selecting $\eta<1/L$ and m large enough to let $\lambda^{\text{SARAH}}<1$ establishes SARAH's linear convergence. For example, choosing $\eta=1/(2L)$ and $m=5\kappa$, we have $\lambda^{\text{SARAH}}\approx 0.8$. □

Proof of Corollary 2

Proof. If we choose $\eta = 1/(2L)$ and $m = 6\kappa = 3/(\mu\eta)$, we have $\delta = 1/(2\kappa)$ and $c \geq 4\kappa$, which implies that

$$(1-\mu\eta)^{\frac{1}{\mu\eta}} \leq 0.4$$

(Actually $(1-\mu\eta)^{\frac{1}{\mu\eta}} \approx 1/e$ when $\mu\eta$ small enough). Using the value of η and m, it can be verified that $\lambda^{\text{SVRG}} \leq 0.75$. This implies that $\mathcal{O}\big(\ln\frac{1}{\epsilon}\big)$ outer loops are needed for an ϵ-accurate solution. And since $m = \mathcal{O}(\kappa)$, the overall IFO complexity is $\mathcal{O}\big((n+\kappa)\ln\frac{1}{\epsilon}\big)$. □

C Proofs for BB-SVRG and BB-SARAH

Derivation of (8)**:** It clearly holds that

$$\eta^s = \frac{1}{\theta_\kappa}\frac{\|\tilde{\mathbf{x}}^{s-1}-\tilde{\mathbf{x}}^{s-2}\|^2}{\big\langle \tilde{\mathbf{x}}^{s-1}-\tilde{\mathbf{x}}^{s-2}, \nabla f(\tilde{\mathbf{x}}^{s-1})-\nabla f(\tilde{\mathbf{x}}^{s-2})\big\rangle} \leq \frac{1}{\theta_\kappa}\frac{\|\tilde{\mathbf{x}}^{s-1}-\tilde{\mathbf{x}}^{s-2}\|^2}{\mu\|\tilde{\mathbf{x}}^{s-1}-\tilde{\mathbf{x}}^{s-2}\|^2} = \frac{1}{\theta_\kappa\mu}$$

where the inequality follows since under Assumption 3 (or 4) $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \geq \mu\|\mathbf{x}-\mathbf{y}\|^2$ [Nesterov, 2004, Theorem 2.1.9]. On the other hand, we have

$$\eta^s \geq \frac{1}{\theta_\kappa}\frac{\|\tilde{\mathbf{x}}^{s-1}-\tilde{\mathbf{x}}^{s-2}\|^2}{\|\tilde{\mathbf{x}}^{s-1}-\tilde{\mathbf{x}}^{s-2}\|\|\nabla f(\tilde{\mathbf{x}}^{s-1})-\nabla f(\tilde{\mathbf{x}}^{s-2})\|} \geq \frac{1}{\theta_\kappa L}$$

where the first inequality follows from the Cauchy-Schwarz inequality; and the second inequality is due to Assumption 1.

C.1 Proof for Proposition 1

For BB-SVRG, the step size η^s changes across different inner loops. Since η^s influences convergence, we will use λ^s to denote the convergence rate of the inner loop s, that is, $\mathbb{E}[f(\tilde{\mathbf{x}}^s) - f(\mathbf{x}^*)] \leq \lambda^s \mathbb{E}[f(\tilde{\mathbf{x}}^{s-1}) - f(\mathbf{x}^*)]$.
BB-SVRG with U-Avg:

Proof. From [Johnson and Zhang, 2013], we have the convergence rate is

$$\lambda^s = \frac{1}{\mu\eta^s(1-2\eta^s L)m} + \frac{2\eta^s L}{1-2\eta^s L} \overset{(a)}{\leq} \frac{\kappa\theta_\kappa}{m(1-2\kappa/\theta_\kappa)} + \frac{2\kappa/\theta_\kappa}{1-2\kappa/\theta_\kappa}$$

where (a) is due to (8). Hence, by choosing $\theta_\kappa > 4\kappa$ with $\theta_\kappa = \mathcal{O}(\kappa)$ and $m = \mathcal{O}(\kappa^2)$ such that $\lambda^s < 1$, and using similar arguments as in the proof of Corollary 1, one can readily verify that the IFO complexity is $\mathcal{O}\big((n+\kappa^2)\ln\frac{1}{\epsilon}\big)$. □

BB-SVRG with W-Avg:

Proof. It follows from Theorem 1 and (8) that the convergence rate satisfies

$$\begin{aligned}\lambda^s &= \frac{1}{1-(1-\mu\eta^s)^{m-1}}\left[\frac{(1-\mu\eta^s)^m}{1-2\eta^s L} + \frac{2\mu L(\eta^s)^2(1-\mu\eta)^{m-1}}{1-2L\eta^s} + \frac{2L\eta^s}{1-2L\eta^s}\right] \\ &\leq \frac{1}{1-\big(1-\frac{1}{\kappa\theta_\kappa}\big)^{m-1}}\left[\frac{\big(1-\frac{1}{\kappa\theta_\kappa}\big)^m}{1-2\kappa/\theta_\kappa} + \frac{\frac{2\kappa}{(\theta_\kappa)^2}\big(1-\frac{1}{\kappa\theta_\kappa}\big)^{m-1}}{1-2\kappa/\theta_\kappa} + \frac{2\kappa/\theta_\kappa}{1-2\kappa/\theta_\kappa}\right]\end{aligned}$$

where the inequality is due to (8). Hence, by choosing $\theta_\kappa > 4\kappa$ with $\theta_\kappa = \mathcal{O}(\kappa)$ and $m = \mathcal{O}(\kappa^2)$ so that $\lambda^s < 1$, and using similar arguments as in the proof of Corollary 1, one can establish that the IFO complexity is $\mathcal{O}\big((n+\kappa^2)\ln\frac{1}{\epsilon}\big)$. □

C.2 Proof for Proposition 2

Also for BB-SARAH, the step size η^s changes across different inner loops. Since here too η^s affects convergence, we will use λ^s to denote the convergence rate of the inner loop s; that is, $\mathbb{E}[\|f(\tilde{\mathbf{x}}^s)\|^2] \leq \lambda^s \mathbb{E}[\|f(\tilde{\mathbf{x}}^{s-1})\|^2]$.

BB-SARAH with U-Avg:

Proof. We have from [Nguyen et al., 2017] that the convergence rate is

$$\lambda^s = \frac{1}{\mu\eta^s m} + \frac{\eta^s L}{2-\eta^s L} \overset{(a)}{\leq} \frac{\kappa\theta_\kappa}{m} + \frac{\kappa/\theta_\kappa}{2-\kappa/\theta_\kappa}$$

where (a) is due to (8). Hence, by choosing $\theta_\kappa > \kappa$ with $\theta_\kappa = \mathcal{O}(\kappa)$ and $m = \mathcal{O}(\kappa^2)$ so that $\lambda^s < 1$, and using arguments similar to those in the proof of Corollary 2, one can establish that the IFO complexity is $\mathcal{O}\big((n+\kappa^2)\ln\frac{1}{\epsilon}\big)$. □

BB-SARAH with L-Avg:

Proof. Since the derivation in [Li et al., 2019] relies on Assumption 3, we will first establish the convergence rate under Assumption 4. The proof proceeds along the lines of [Li et al., 2019], except for the use of Lemma 5 to bound $\mathbb{E}[\|\mathbf{v}_t^s\|]^2$. After a simple derivation, one can have the convergence rate

$$\lambda^s = \frac{2\eta^s L}{2-\eta^s L} + 2(1+\eta^s L)\bigg(1-\frac{2\eta^s L}{1+\kappa}\bigg)^m.$$

Then using (8) to upper bound λ^s, we have

$$\lambda^s \leq \frac{2\kappa/\theta_\kappa}{2-\kappa/\theta_\kappa} + 2(1+\kappa/\theta_\kappa)\bigg(1-\frac{2}{(1+\kappa)\theta_\kappa}\bigg)^m.$$

Hence, by choosing $\theta_\kappa > 3\kappa/2$ with $\theta_\kappa = \mathcal{O}(\kappa)$ and $m = \mathcal{O}(\kappa^2)$ so that $\lambda^s < 1$, and using arguments similar to those in the proof of Corollary 2, one can verify that the IFO complexity is $\mathcal{O}\big((n+\kappa^2)\ln\frac{1}{\epsilon}\big)$. □

BB-SARAH with W-Avg:

Proof. From Theorem 2, the convergence rate is

$$\begin{aligned}\lambda^s &= \frac{(1-\mu\eta^s)^m}{c\mu\eta^s} + \bigg[(1-\mu\eta^s)^m - \Big(1-\frac{2\eta^s L}{1+\kappa}\Big)^m\bigg]\frac{L+\mu}{c(L-\mu)} + \frac{\eta^s L(m-1)}{c(2-\eta^s L)} + \frac{2-2\eta^s L}{2-\eta^s L}\frac{1+\kappa}{2c\eta^s L}\\ &\leq \frac{\kappa\theta_\kappa\big(1-\frac{1}{\kappa\theta_\kappa}\big)^m}{c} + \big(1-\frac{1}{\kappa\theta_\kappa}\big)^m\frac{L+\mu}{c(L-\mu)} + \frac{(m-1)\kappa/\theta_\kappa}{c(2-\kappa/\theta_\kappa)} + \frac{2}{2-\kappa/\theta_\kappa}\frac{(1+\kappa)\theta_\kappa}{2c}\end{aligned}$$

where $c = m - \frac{1}{\mu\eta^s} + \frac{(1-\mu\eta^s)^m}{\mu\eta^s} \geq m - \frac{1}{\mu\eta^s} \geq m - \kappa\theta_\kappa$. With $\theta_\kappa = \mathcal{O}(\kappa)$ and $m = \mathcal{O}(\kappa^2)$ so that $c = \mathcal{O}(\kappa^2)$, we find that $\lambda^s < 1$. In addition, since $\eta^s < 1/L$ is still needed to guarantee convergence (cf. Theorem 2), one must have $\theta_\kappa > \kappa$. □

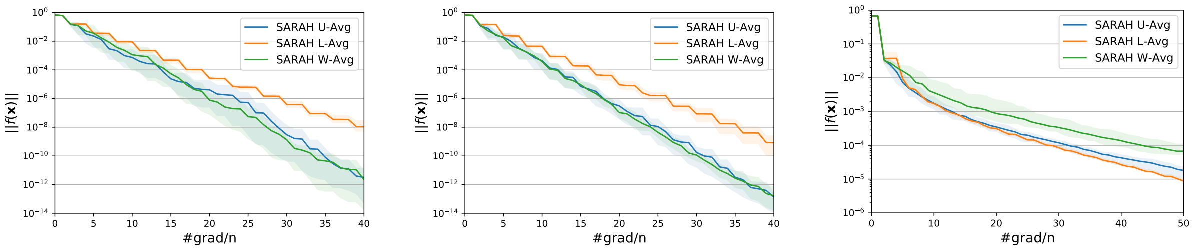
D More on Numerical Experiments

D.1 More Numerical Tests of Section 3.4

This subsection presents additional numerical tests to support that averaging is not merely a 'proof trick.' Specifically, experiments with SARAH under different types of averaging on dataset *a9a* are showcased in Fig. 6. Similar to the performance of SARAH on dataset *w7a*, W-Avg is better when the step size is chosen large, while a smaller step size favors L-Avg.

D.2 Details of datasets used in Section 5

The dimension d, number of training data n, the weight used for regularization, and other details of datasets used in Section 5, are listed in Table 1.



(a) $\eta = 0.9/L$ (a) $\eta = 0.6/L$ (b) $\eta = 0.06/L$

Figure 6: Comparing SARAH with different types of averaging on dataset *a9a*. In all tests, we set $\mu = 0.002$ with $m = 5\kappa$.

Table 1: Parameters of datasets used in numerical tests

Dataset	d	n (train)	density	n (test)	μ
a9a	122	$3,185$	11.37%	$29,376$	0.001
rcv1	$47,236$	$20,242$	0.157%	$677,399$	0.00025
real-sim	$20,958$	$50,617$	0.24%	$21,692$	0.00025