Estimate Sequences for Variance-Reduced Stochastic Composite Optimization

Andrei Kulunchakov, Julien Mairal

Estimate Sequences for Variance-Reduced Stochastic Composite Optimization

Andrei Kulunchakov [1] **Julien Mairal** [1]

Abstract

In this paper, we propose a unified view of gradient-based algorithms for stochastic convex composite optimization by extending the concept of estimate sequence introduced by Nesterov. This point of view covers the stochastic gradient descent method, variants of the approaches SAGA, SVRG, and has several advantages: (i) we provide a generic proof of convergence for the aforementioned methods; (ii) we show that this SVRG variant is adaptive to strong convexity; (iii) we naturally obtain new algorithms with the same guarantees; (iv) we derive generic strategies to make these algorithms robust to stochastic noise, which is useful when data is corrupted by small random perturbations. Finally, we show that this viewpoint is useful to obtain new accelerated algorithms in the sense of Nesterov.

1. Introduction

We consider convex optimization problems of the form

$$\min_{x \in \mathbb{R}^p} \left\{ F(x) := f(x) + \psi(x) \right\}, \tag{1}$$

where f is μ-strongly convex and L-smooth (differentiable with L-Lipschitz continuous gradient), and ψ is convex lower-semicontinuous. For instance, ψ may be the ℓ_1-norm but may also be the indicator function of a convex set $\mathcal{C}$ for constrained problems (Hiriart-Urruty & Lemaréchal, 1996).

More specifically, we focus on stochastic objectives, where f is an expectation or a finite sum of convex functions

$$f(x) = \mathbb{E}_\xi \left[\tilde{f}(x, \xi) \right] \quad \text{or} \quad f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x). \tag{2}$$

On the left, ξ is a random variable representing a data point drawn according to some distribution and $\tilde{f}(x,\xi)$ measures the fit of some model parameter x to the data point ξ.

[1]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France. Correspondence to: Julien Mairal <julien.mairal@inria.fr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

While the finite-sum setting is a particular case of expectation, the deterministic nature of the resulting cost function drastically changes the performance guarantees an optimization method may achieve to solve (1). In particular, when an algorithm is only allowed to access unbiased measurements of the objective and gradient, it may be shown that the worst-case convergence rate in expected function value cannot be better than $O(1/k)$ in general, where k is the number of iterations (Nemirovski et al., 2009; Agarwal et al., 2012).

Even though this pessimistic result applies to the general stochastic case, linear convergence rates can be obtained for deterministic finite sums. For instance, linear rates are achieved by SAG (Schmidt et al., 2017), SAGA (Defazio et al., 2014), SVRG (Johnson & Zhang, 2013; Xiao & Zhang, 2014), SDCA (Shalev-Shwartz & Zhang, 2016), MISO (Mairal, 2015), Katyusha (Allen-Zhu, 2017), MiG (Zhou et al., 2018), SARAH (Nguyen et al., 2017), accelerated SAGA (Zhou, 2019) or Lan & Zhou (2018a). In the non-convex case, a recent focus has been on improving convergence rates for finding first-order stationary points (Lei et al., 2017; Paquette et al., 2018; Fang et al., 2018), which is however beyond the scope of our paper. A common interpretation is to see these algorithms as performing SGD steps with an estimate of the gradient that has lower variance.

In this paper, we are interested in providing a unified view of such stochastic optimization algorithms, but we also want to investigate their *robustness* to random pertubations. Specifically, we consider objective functions with an explicit finite-sum structure such as (2) when only noisy estimates of the gradients $\nabla f_i(x)$ are available. Such a setting may occur for various reasons. Perturbations may be injected during training to achieve better generalization (Srivastava et al., 2014), perform stable feature selection (Meinshausen & Bühlmann, 2010), for privacy-aware learning (Wainwright et al., 2012) or to improve model robustness (Zheng et al., 2016).

Each point indexed by i is then corrupted by a random perturbation ρ_i and the function f may be written as

$$f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x) \quad \text{with} \quad f_i(x) = \mathbb{E}_{\rho_i} \left[\tilde{f}_i(x, \rho_i) \right]. \tag{3}$$

Since the exact gradients $\nabla f_i(x)$ cannot be computed, all the aforementioned variance-reduction methods do not apply and the standard approach is to use SGD. Typically, the variance of the gradient estimate then decomposes into two parts $\sigma^2 = \sigma_s^2 + \tilde{\sigma}^2$, where σ_s^2 is due to data sampling and $\tilde{\sigma}^2$ to the random perturbation. In such a context, variance reduction consists of designing algorithms whose convergence rate depends on $\tilde{\sigma}^2$, which is potentially much smaller than σ^2. The SAGA and SVRG methods have been adapted for such a purpose by Hofmann et al. (2015a), though the resulting algorithms have non-zero asymptotic error; the MISO method was adapted by Bietti & Mairal (2017) at the cost of a memory overhead of $O(np)$, whereas other variants of SAGA and SVRG have been proposed by Zheng & Kwok (2018) for linear models in machine learning.

In this paper, we extend estimate sequences introduced by Nesterov (2004), which are typically used to design accelerated algorithms. Estimate sequences have been used before for stochastic optimization (Devolder, 2011; Lin et al., 2014; Lu & Xiao, 2015), but not for the following generic purpose:

First, we make a large class of variance-reduction methods robust to stochastic perturbations. More precisely, by using a sampling strategy Q to select indices of the sum (3) at each iteration, when each function f_i is convex and L_i-smooth, the worst-case iteration complexity of our approaches in function values—that is, the number of iterations to guarantee $\mathbb{E}[F(x_k) - F^*] \leq \varepsilon$—is upper bounded by

$$O\left(\left(n + \frac{L_Q}{\mu}\right)\log\left(\frac{F(x_0)-F^*}{\varepsilon}\right)\right) + O\left(\frac{\rho_Q\tilde{\sigma}^2}{\mu\varepsilon}\right),$$

where $\rho_Q \geq 1$ and L_Q depends on Q. For the uniform distribution, we have $\rho_Q = 1$ and $L_Q = \max_i L_i$, whereas a non-uniform Q may yield $L_Q = \frac{1}{n}\sum_i L_i$. The term on the left corresponds to the complexity of variance-reduction methods without perturbation, and $O(\tilde{\sigma}^2/\mu\varepsilon)$ is the optimal sublinear rate of convergence for a stochastic optimization problem when the gradient estimates have variance $\tilde{\sigma}^2$. In contrast, a variant of SGD applied to (3) has worst-case complexity $O(\sigma^2/\mu\varepsilon)$, with potentially $\sigma^2 \gg \tilde{\sigma}^2$.

Second, we design a new accelerated algorithm which, to our knowledge, is the first one to achieve the complexity

$$O\left(\left(n + \sqrt{n\frac{L_Q}{\mu}}\right)\log\left(\frac{F(x_0)-F^*}{\varepsilon}\right)\right) + O\left(\frac{\rho_Q\tilde{\sigma}^2}{\mu\varepsilon}\right),$$

where the term on the left matches the optimal complexity for finite sums when $\tilde{\sigma}^2 = 0$ (Arjevani & Shamir, 2016), which has been achieved by Allen-Zhu (2017); Zhou et al. (2018); Zhou (2019); Kovalev et al. (2019). The general stochastic finite-sum problem with $\tilde{\sigma}^2 > 0$ was also considered with an acceleration mechanism by Lan & Zhou (2018b) for distributed optimization being optimal in terms of communication rounds, but not in the global complexity.

Note that we treat here only the strongly convex case, but similar results can be obtained when $\mu = 0$, as shown in a long version of this paper (Kulunchakov & Mairal, 2019).

2. A Generic Framework

In this section, we introduce stochastic estimate sequences and show how they can handle variance reduction.

2.1. A Classical Iteration Revisited

Consider an algorithm that performs the following updates:

$$x_k \leftarrow \text{Prox}_{\eta_k\psi}\left[x_{k-1} - \eta_k g_k\right], \qquad \text{(A)}$$

where $\mathbb{E}[g_k|\mathcal{F}_{k-1}] = \nabla f(x_{k-1})$ and $\mathcal{F}_{k-1}$ is the filtration representing all information up to iteration k–1, $\eta_k > 0$ is a step size, and $\text{Prox}_{\eta\psi}[.]$ is the proximal operator (Moreau, 1962) defined for any scalar $\eta > 0$ as the unique solution of

$$\text{Prox}_{\eta\psi}[u] := \operatorname*{argmin}_{x\in\mathbb{R}^p}\left\{\eta\psi(x) + \frac{1}{2}\|x-u\|^2\right\}. \qquad (4)$$

Key to our analysis, we interpret (A) as the iterative minimization of quadratic surrogate functions.

Interpretation with stochastic estimate sequence. Consider

$$d_0(x) = d_0^* + \frac{\gamma_0}{2}\|x - x_0\|^2, \qquad (5)$$

with $\gamma_0 \geq \mu$ and d_0^* is a scalar value that is left unspecified at the moment. Then, it is easy to show that x_k in (A) minimizes d_k defined recursively for $k \geq 1$ as

$$d_k(x) = (1-\delta_k)d_{k-1}(x) + \delta_k l_k(x), \qquad (6)$$

where

$$\begin{aligned} l_k(x) &= f(x_{k-1}) + g_k^\top(x - x_{k-1}) \\ &\quad + \frac{\mu}{2}\|x - x_{k-1}\|^2 + \psi(x_k) + \psi'(x_k)^\top(x - x_k), \end{aligned}$$

δ_k, γ_k satisfy the system of equations

$$\delta_k = \eta_k\gamma_k \quad \text{and} \quad \gamma_k = (1-\delta_k)\gamma_{k-1} + \mu\delta_k, \qquad (7)$$

(note that $\gamma_0 = \mu$ yields $\gamma_k = \mu$ for all k), and

$$\psi'(x_k) = \frac{1}{\eta_k}(x_{k-1} - x_k) - g_k.$$

Here, $\psi'(x_k)$ is a subgradient in $\partial\psi(x_k)$. By simply using the definition of the proximal operator (4) and considering first-order optimality conditions, we indeed have that $0 \in x_k - x_{k-1} + \eta_k g_k + \eta_k\partial\psi(x_k)$ and x_k coincides with the minimizer of d_k. This allows us to write d_k in the form

$$d_k(x) = d_k^* + \frac{\gamma_k}{2}\|x - x_k\|^2 \quad \text{for all } k \geq 0.$$

The construction (6) is akin to that of estimate sequences introduced by Nesterov (2004), which are typically used for designing accelerated gradient-based optimization algorithms. In this section, we are however not interested in acceleration, but instead in stochastic optimization and variance reduction. One of the main properties of estimate sequences that we will nevertheless use is their ability to behave asymptotically as a lower bound. Indeed, we have

$$\mathbb{E}[d_k(x^*)] \leq \Gamma_k d_0(x^*) + (1 - \Gamma_k)F^*, \tag{8}$$

where x^* is a minimizer of F and $\Gamma_k = \prod_{t=1}^{k}(1-\delta_t)$. The inequality comes from strong convexity since $\mathbb{E}[g_k^\top(x^* - x_{k\text{-}1})|\mathcal{F}_{k\text{-}1}] = \nabla f(x_{k\text{-}1})^\top(x^* - x_{k\text{-}1})$, leading to the relation $\mathbb{E}[d_k(x^*)] \leq (1-\delta_k)\mathbb{E}[d_{k\text{-}1}(x^*)] + \delta_k F^*$. Then, by unrolling the recursion, we obtain (8). When Γ_k converges to zero, the contribution of the initial surrogate d_0 disappears and $\mathbb{E}[d_k(x^*)]$ behaves as a lower bound of F^*.

Relation with existing algorithms. The iteration (A) encompasses many approaches such as ISTA (proximal gradient uses the exact gradient $g_k = \nabla f(x_{k\text{-}1})$ (Beck & Teboulle, 2009; Nesterov, 2013) or proximal variants of the stochastic gradient descent method to deal with a composite objective (Lan, 2012). Of interest to us, the variance-reduced stochastic optimization approaches SVRG (Xiao & Zhang, 2014) and SAGA (Defazio et al., 2014) also follow (A) but with an unbiased gradient estimator g_k whose variance reduces over time. Specifically, they use

$$g_k = \nabla f_{i_k}(x_{k\text{-}1}) - z_{k\text{-}1}^{i_k} + \bar{z}_{k\text{-}1} \quad \text{with} \quad \bar{z}_{k\text{-}1} = \frac{1}{n}\sum_{i=1}^{n} z_{k\text{-}1}^i, \tag{9}$$

where i_k is an index chosen uniformly in $\{1,\ldots,n\}$ at random, and each z_k^i is equal to a gradient $\nabla f_i(\tilde{x}_k^i)$, where $\tilde{x}_k^i$ is one of the previous iterates. The motivation is that given two random variables X and Y, it is possible to define a new variable $Z = X - Y + \mathbb{E}[Y]$ which has the same expectation as X but potentially a lower variance if Y is positively correlated with X. SVRG uses the same anchor point $\tilde{x}_k^i = \tilde{x}_k$ for all i, where $\tilde{x}_k$ is updated every m iterations. Typically, the memory cost of SVRG is that of storing the variable $\tilde{x}_k$ and the gradient $\bar{z}_k = \nabla f(\tilde{x}_k)$, which is thus $O(p)$. On the other hand, SAGA updates only $z_k^{i_k} = \nabla f_{i_k}(x_{k\text{-}1})$ at iteration k, such that $z_k^i = z_{k\text{-}1}^i$ if $i \neq i_k$. Thus, SAGA requires storing n gradients. While in general the overhead cost in memory is of order $O(np)$, it may be reduced to $O(n)$ when dealing with linear models in machine learning (see Defazio et al., 2014). Note that variants with non-uniform sampling of the indices i_k have been proposed by Xiao & Zhang (2014); Schmidt et al. (2015), which we discuss later.

In order to make our proofs consistent for SAGA and SVRG (and MISO in Kulunchakov & Mairal, 2019), we consider a variant of SVRG with a randomized gradient updating schedule (Hofmann et al., 2015a). Remarkably, this variant was shown to provide benefits over the fixed schedule in a concurrent work (Kovalev et al., 2019) when $\tilde{\sigma}^2 = 0$.

2.2. Gradient Estimators and New Algorithms

In this paper, we consider the gradient estimators below. For all of them, we define the variance σ_k to be

$$\sigma_k^2 = \mathbb{E}\left[\|g_k - \nabla f(x_{k\text{-}1})\|^2\right].$$

ISTA. Simply consider $g_k = \nabla f(x_{k\text{-}1})$ and $\sigma_k = 0$.

SGD. We assume that g_k has variance bounded by σ^2. Typically, when $f(x) = \mathbb{E}_\xi[\tilde{f}(x,\xi)]$, a data point ξ_k is drawn at iteration k and $g_k = \nabla\tilde{f}(x,\xi_k)$. Even though the bounded variance assumption has limitations, it remains the most standard one for stochastic optimization and more realistic settings (such as (Bottou et al., 2018; Nguyen et al., 2018) for the smooth case) are left for future work.

random-SVRG. For finite sums, we consider a variant of SVRG with random update of the anchor point $\tilde{x}_{k\text{-}1}$, proposed originally in (Hofmann et al., 2015b), combined with non-uniform sampling. Specifically, g_k is defined as

$$g_k = \frac{1}{q_{i_k} n}\left(\tilde{\nabla} f_{i_k}(x_{k\text{-}1}) - z_{k\text{-}1}^{i_k}\right) + \bar{z}_{k\text{-}1}, \tag{10}$$

where $i_k \sim Q = \{q_1,\ldots,q_n\}$ and $\tilde{\nabla}$ denotes a perturbed gradient operator. For instance, if $f_i(x) = \mathbb{E}_\rho[\tilde{f}_i(x,\rho)]$ for all i, where ρ is a stochastic perturbation, instead of accessing $\nabla f_{i_k}(x_{k\text{-}1})$, we draw a perturbation ρ_k and observe

$$\tilde{\nabla} f_{i_k}(x_{k\text{-}1}) = \nabla\tilde{f}_{i_k}(x_{k\text{-}1},\rho_k) = \nabla f_{i_k}(x_{k\text{-}1}) + \zeta_k,$$

where the perturbation ζ_k has zero mean given $\mathcal{F}_{k\text{-}1}$ and its variance is bounded by $\tilde{\sigma}^2$. When considering the setting without perturbation, we simply have $\tilde{\nabla} = \nabla$.

Similar to the previous case, the variables z_k^i and $\bar{z}_k$ also correspond to noisy estimates of the gradients. Specifically,

$$z_k^i = \tilde{\nabla} f_i(\tilde{x}_k) \quad \text{and} \quad \bar{z}_k = \frac{1}{n}\sum_{i=1}^{n} z_k^i,$$

where $\tilde{x}_k$ is an anchor point that is updated on average every n iterations. Whereas the classical SVRG approach updates $\tilde{x}_k$ on a fixed schedule, we perform random updates: with probability $1/n$, we choose $\tilde{x}_k = x_k$ and recompute $\bar{z}_k = \tilde{\nabla} f(\tilde{x}_k)$; otherwise $\tilde{x}_k$ is kept unchanged. In comparison with the fixed schedule, the analysis with the random one is simplified and can be unified with that of SAGA. This approach is described in Algorithm 1.

In terms of memory, the random-SVRG gradient estimator requires to store an anchor point $\tilde{x}_{k\text{-}1}$ and the average gradients $\bar{z}_{k\text{-}1}$. The z_k^i's do not need to be stored; only the n random seeds to produce the perturbations are kept into memory, which allows us to compute $z_{k\text{-}1}^{i_k} = \tilde{\nabla} f_{i_k}(\tilde{x}_{k\text{-}1})$ at iteration k, with the same perturbation for index i_k that was used to compute $\bar{z}_{k\text{-}1} = \frac{1}{n}\sum_{i=1}^{n} z_{k\text{-}1}^i$ when the anchor point was last updated. The overall cost is thus $O(n+p)$.

Algorithm 1 Iteration (A) with random-SVRG estimator

Input: x_0 in $\mathbb{R}^p$ (initial point); K (number of iterations); $(\eta_k)_{k\geq 0}$ (step sizes); $\gamma_0 \geq \mu$ (if averaging);
Initialization: $\tilde{x}_0 = \hat{x}_0 = x_0$; $\bar{z}_0 = \frac{1}{n}\sum_{i=1}^n \tilde{\nabla} f_i(\tilde{x}_0)$;
for $k = 1, \ldots, K$ **do**
  Sample i_k according to $Q = \{q_1, \ldots, q_n\}$;
  Compute the gradient estimator with perturbations:
$$g_k = \frac{1}{q_{i_k} n}\left(\tilde{\nabla} f_{i_k}(x_{k\text{-}1}) - \tilde{\nabla} f_{i_k}(\tilde{x}_{k\text{-}1})\right) + \bar{z}_{k\text{-}1};$$
  Compute the next iterate
$$x_k \leftarrow \text{Prox}_{\eta_k \psi}\left[x_{k\text{-}1} - \eta_k g_k\right];$$
  With probability $1/n$,
$$\tilde{x}_k = x_k \quad \text{and} \quad \bar{z}_k = \frac{1}{n}\sum_{i=1}^n \tilde{\nabla} f_i(\tilde{x}_k);$$
  Otherwise, with probability $1 - 1/n$, keep the anchor point unchanged $\tilde{x}_k = \tilde{x}_{k\text{-}1}$ and $\bar{z}_k = \bar{z}_{k\text{-}1}$;
  Optional: Use online averaging using δ_k from (7):
$$\hat{x}_k = (1-\tau_k)\hat{x}_{k\text{-}1} + \tau_k x_k \ \text{ with } \tau_k = \min\left(\delta_k, \frac{1}{5n}\right);$$
end for
Output: x_K or $\hat{x}_K$ if averaging.

SAGA. The estimator has a form similar to (10) but with a different choice of variables z_k^i. Unlike SVRG that stores an anchor point $\tilde{x}_k$, the SAGA estimator requires storing and incrementally updating the n auxiliary variables z_k^i, while maintaining the relation $\bar{z}_k = \frac{1}{n}\sum_{i=1}^n z_k^i$. We consider variants such that each gradient $\nabla f_i(x)$ is corrupted by a random perturbation; to deal with non-uniform sampling, we use a similar strategy as Schmidt et al. (2015). The corresponding algorithm is available in Appendix A.

3. Convergence Analysis and Robustness

In Section 3.1, we present a general convergence result and the analysis with variance-reduction is presented in Section 3.2. All proofs are in the appendix.

3.1. Generic Convergence Result

The following proposition gives a key relation between $F(x_k)$, the surrogate d_k, $d_{k\text{-}1}$ and the variance σ_k.

Proposition 1 (**Key relation).** *For iteration (A), assuming $\eta_k \leq 1/L$, we have for all $k \geq 1$,*

$$\begin{aligned}&\delta_k(\mathbb{E}[F(x_k)] - F^*) + \mathbb{E}[d_k(x^*) - d_k^*] \\ &\qquad \leq (1-\delta_k)\mathbb{E}[d_{k\text{-}1}(x^*) - d_{k\text{-}1}^*] + \eta_k\delta_k\sigma_k^2, \quad (11)\end{aligned}$$

where x^* is a minimizer of F and $F^* = F(x^*)$.

Then, without making further assumption on σ_k, we have the following general convergence result, which is a direct consequence of the averaging Lemma C.8 in the appendix, inspired in part by Ghadimi & Lan (2012):

Theorem 1 (**General convergence result).** *Under the same assumptions as in Proposition 1, by using the averaging strategy of Lemma C.8, which produces an iterate $\hat{x}_k$,*

$$\begin{aligned}&\mathbb{E}\left[F(\hat{x}_k) - F^* + \frac{\gamma_k}{2}\|x_k - x^*\|^2\right] \\ &\leq \Gamma_k\left(F(x_0) - F^* + \frac{\gamma_0}{2}\|x_0 - x^*\|^2 + \sum_{t=1}^k \frac{\delta_t\eta_t\sigma_t^2}{\Gamma_t}\right),\end{aligned}$$

where $\Gamma_k = \prod_{t=1}^k(1-\delta_t)$ and x^* is a minimizer of F.

Theorem 1 allows us to recover convergence rates for various algorithms. In the corollary below, we consider the stochastic setting with constant step sizes; the algorithm converges with the same rate as the deterministic problem to a noise-dominated region of radius σ^2/L. The proof simply uses Lemma C.6, which provides the convergence rate of $(\Gamma_k)_{k\geq 0}$ and uses the relation $\Gamma_k\sum_{t=1}^k \frac{\delta_t}{\Gamma_t} = 1 - \Gamma_k \leq 1$ from Lemma C.5 in the appendix.

Corollary 1 (**Prox-SGD with constant step-size).** *Assume in Theorem 1 that $\sigma_k \leq \sigma$, and choose $\gamma_0 = \mu$ and $\eta_k = 1/L$. Then,*

$$\begin{aligned}&\mathbb{E}\left[F(\hat{x}_k) - F^* + \frac{\mu}{2}\|x_k - x^*\|^2\right] \\ &\leq \left(1 - \frac{\mu}{L}\right)^k\left(F(x_0) - F^* + \frac{\mu}{2}\|x_0 - x^*\|^2\right) + \frac{\sigma^2}{L}. \quad (12)\end{aligned}$$

We now show that it is also possible to obtain converging algorithms by using decreasing step sizes.

Corollary 2 (**Prox-SGD with decreasing step-sizes).** *Assume that we target an accuracy ε smaller than $2\sigma^2/L$. First, use iteration (A) as in Theorem 1 with a constant step-size $\eta_k = 1/L$ and $\gamma_0 = \mu$, leading to the convergence rate (12), until $\mathbb{E}[F(\hat{x}_k) - F^*] \leq 2\sigma^2/L$. Then, restart the optimization procedure with decreasing step-sizes $\eta_k = \min\left(\frac{1}{L}, \frac{2}{\mu(k+2)}\right)$. The resulting number of gradient evaluations to achieve $\mathbb{E}[F(\hat{x}_k) - F^*] \leq \varepsilon$ is upper bounded by*

$$O\left(\frac{L}{\mu}\log\left(\frac{F(x_0) - F^*}{\varepsilon}\right)\right) + O\left(\frac{\sigma^2}{\mu\varepsilon}\right).$$

We note that the dependency in σ^2 with the rate $O(\sigma^2/\mu\varepsilon)$ is optimal for strongly convex functions (Nemirovski et al., 2009). Unfortunately, estimating σ is not easy and knowing in practice when to start decreasing the step sizes in SGD algorithms is an open problem. The corollary simply supports the heuristic consisting of adopting a constant step size long enough until the iterates oscillate without much progress, before decreasing the step sizes (see Bottou et al., 2018).

3.2. Faster Convergence with Variance Reduction

Stochastic variance-reduced algorithms rely on gradient estimates whose variance decreases as fast as the objective. Our framework provides a unified proof of convergence for our variants of SVRG and SAGA and makes them robust to stochastic perturbations. Specifically, we consider the minimization of a finite sum of functions as in (3), but each observation of the gradient $\nabla f_i(x)$ is corrupted by a random noise variable. The next proposition extends the proof of SVRG (Xiao & Zhang, 2014) and characterizes σ_k^2.

Proposition 2 (Generic Upper-Bound on Variance). *Consider the optimization problem (1) when f is a finite sum of functions $f = \frac{1}{n}\sum_{i=1}^n f_i$ where each f_i is convex and L_i-smooth with $L_i \geq \mu$. Then, the random-SVRG and SAGA gradient estimates defined in Section 2.2 satisfy*

$$\sigma_k^2 \leq 4L_Q \mathbb{E}[F(x_{k\text{-}1}) - F^*] + \frac{2}{n}\mathbb{E}\left[\sum_{i=1}^n \frac{1}{nq_i}\|u_{k\text{-}1}^i - \nabla f_i(x^*)\|^2\right] + 3\rho_Q\tilde{\sigma}^2, \quad (13)$$

where $L_Q = \max_i L_i/(q_i n)$, $\rho_Q = 1/(n\min_i q_i)$, and for all i and k, u_k^i is equal to z_k^i without noise—that is

$$u_k^i = \nabla f_i(\tilde{x}_k) \quad \text{for random-SVRG}$$
$$u_k^{j_k} = \nabla f_{j_k}(x_k) \quad \text{and} \quad u_k^j = u_{k\text{-}1}^j \quad \text{if} \quad j \neq j_k \quad \text{for SAGA}.$$

Next, we apply this result to Proposition 1.

Proposition 3 (Lyapunov function). *Consider the same setting as Proposition 2 and the same gradient estimators. When using the construction of d_k from Sections 2.1, and assuming $\gamma_0 \geq \mu$ and $(\eta_k)_{k\geq 0}$ is non-increasing with $\eta_k \leq \frac{1}{12L_Q}$, we have for all $k \geq 1$, with $\tau_k = \min\left(\delta_k, \frac{1}{5n}\right)$,*

$$\frac{\delta_k}{6}\mathbb{E}[F(x_k) - F^*] + T_k \leq (1-\tau_k)\,T_{k\text{-}1} + 3\rho_Q\eta_k\delta_k\tilde{\sigma}^2,$$
$$\text{where } T_k = 5L_Q\eta_k\delta_k\mathbb{E}[F(x_k) - F^*] + \mathbb{E}[d_k(x^*) - d_k^*] + \frac{5\eta_k\delta_k}{2}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \frac{1}{q_i n}\|u_k^i - u_*^i\|^2\right]. \quad (14)$$

From the Lyapunov function, we obtain a general convergence result for the variance-reduced stochastic algorithms.

Theorem 2 (Convergence with variance-reduction). *Consider the same setting as Proposition 3. Then, by using the averaging strategy described in Algorithm 1,*

$$\mathbb{E}\left[F(\hat{x}_k) - F^* + \frac{6\tau_k}{\delta_k}T_k\right] \leq \Theta_k\left(F(x_0) - F^* + \frac{6\tau_k}{\delta_k}T_0 + \frac{18\rho_Q\tau_k\tilde{\sigma}^2}{\delta_k}\sum_{t=1}^k \frac{\eta_t\delta_t}{\Theta_t}\right),$$

where $\Theta_k = \prod_{t=1}^k (1-\tau_t)$.

The theorem is a direct application of the averaging Lemma C.8 to Proposition 3. From this generic convergence theorem, we now study particular cases.

Corollary 3 (Variance-reduction with constant η). *Consider the same setting as in Theorem 2, with $\gamma_0 = \mu$, $\eta_k = \frac{1}{12L_Q}$, and $\tau_k = \tau = \min\left(\frac{\mu}{12L_Q}, \frac{1}{5n}\right)$. Then,*

$$\mathbb{E}\left[F(\hat{x}_k) - F^* + 36L_Q\tau\|x_k - x^*\|^2\right] \leq 8\Theta_k\left(F(x_0) - F^*\right) + \frac{3\rho_Q\tilde{\sigma}^2}{2L_Q}.$$

This first corollary shows that the algorithm achieves a linear convergence rate to a noise-dominated region. Interestingly, the algorithm *without averaging* does not require computing τ and produces iterates $(x_k)_{k\geq 0}$ without using the strong convexity constant μ. *This shows that all estimators we consider can become adaptive to μ.*

Moreover, we note that the non-uniform strategy slightly degrades the dependency in $\tilde{\sigma}^2$: indeed, $\rho_Q = 1$ and $L_Q/\rho_Q = \max_i L_i$ if Q is uniform, but with non-uniform $q_i = L_i/\sum_{j=1}^n L_j$, we have instead $L_Q = \frac{1}{n}\sum_{i=1}^n L_i$ (which is better) and $L_Q/\rho_Q = \min_i L_i$ (which is worse).

Corollary 4 (Variance-reduction with decreasing η_k). *Consider the same setting as in Theorem 2 and target an accuracy $\varepsilon \leq 24\rho_Q\eta\tilde{\sigma}^2$, with $\eta = \min\left(\frac{1}{12L_Q}, \frac{1}{5\mu n}\right)$. Then, use a constant step-size strategy $\eta_k = \eta$ with $\gamma_0 = \mu$ until we find a point $\hat{x}_k$ such that $\mathbb{E}[F(\hat{x}_k) - F^*] \leq 24\rho_Q\eta\tilde{\sigma}^2$. Then, restart the optimization with decreasing step-sizes $\eta_k = \min\left(\eta, \frac{2}{\mu(k+2)}\right)$. The number of gradient evaluations to achieve $\mathbb{E}[F(\hat{x}_k) - F^*] \leq \varepsilon$ is upper bounded by*

$$O\left(\left(n + \frac{L_Q}{\mu}\right)\log\left(\frac{F(x_0) - F^*}{\varepsilon}\right)\right) + O\left(\frac{\rho_Q\tilde{\sigma}^2}{\mu\varepsilon}\right).$$

The corollary shows that variance-reduction algorithms may exhibit an optimal dependency on the noise level $\tilde{\sigma}^2$.

4. Accelerated Stochastic Algorithms

We now consider the following iteration, involving an extrapolation sequence $(y_k)_{k\geq 1}$, which is a classical mechanism from accelerated first-order algorithms (Beck & Teboulle, 2009; Nesterov, 2013). Given a sequence of step-sizes $(\eta_k)_{k\geq 0}$ with $\eta_k \leq 1/L$ for all $k \geq 0$, and $\gamma_0 \geq \mu$, we consider the sequences $(\delta_k)_{k\geq 0}$ and $(\gamma_k)_{k\geq 0}$ that satisfy

$$\delta_k = \sqrt{\eta_k \gamma_k} \quad \text{for all } k \geq 0$$
$$\gamma_k = (1-\delta_k)\gamma_{k-1} + \delta_k \mu \quad \text{for all } k \geq 1.$$

Then, for $k \geq 1$, we consider the iteration

$$x_k = \text{Prox}_{\eta_k \psi}\left[y_{k-1} - \eta_k g_k\right]$$
$$y_k = x_k + \beta_k(x_k - x_{k-1}) \ \text{ with } \beta_k = \frac{\delta_k(1-\delta_k)\eta_{k+1}}{\eta_k \delta_{k+1} + \eta_{k+1}\delta_k^2}, \qquad \text{(B)}$$

where $\mathbb{E}[g_k|\mathcal{F}_{k-1}] = \nabla f(y_{k-1})$. Iteration (B) resembles the accelerated SGD approaches of Hu et al. (2009); Ghadimi & Lan (2012); Lin et al. (2014) but is slightly simpler since it involves two sequences of variables instead of three.

4.1. Convergence Analysis without Variance Reduction

Consider then the stochastic estimate sequence d_k introduced in (6) with d_0 defined as in (5) and

$$l_k(x) = f(y_{k-1}) + g_k^\top(x - y_{k-1}) + \frac{\mu}{2}\|x - y_{k-1}\|^2 + \psi(x_k) + \psi'(x_k)^\top(x - x_k), \quad (15)$$

and $\psi'(x_k) = \frac{1}{\eta_k}(y_{k-1} - x_k) - g_k$ is in $\partial\psi(x_k)$ by definition of the proximal operator. As in Section 2, $d_k(x^*)$ asymptotically becomes a lower bound on F^* since (8) remains satisfied. This time, the iterate x_k does not minimize d_k, and we denote by v_k instead its minimizer, allowing us to write d_k in the canonical form

$$d_k(x) = d_k^* + \frac{\gamma_k}{2}\|x - v_k\|^2.$$

The first lemma highlights classical relations between the iterates $(x_k)_{k\geq 0}$, $(y_k)_{k\geq 0}$ and the minimizers $(v_k)_{k\geq 0}$, which also appears in (Nesterov, 2004, p. 78) for constant η_k. Note that the construction of stochastic estimate sequence resembles that of (Devolder, 2011; Lin et al., 2014). The main difference lies in the choice of function l_k in (15), which yields a different algorithm and slightly stronger guarantees.

Lemma 1 (**Relations between y_k and x_k**). *The sequences $(x_k)_{k\geq 0}$ and $(y_k)_{k\geq 0}$ produced by iteration (B) satisfy for all $k \geq 0$, with $v_0 = y_0 = x_0$,*

$$y_k = \theta_k x_k + (1-\theta_k)v_k \quad \text{with} \quad \theta_k = \frac{\gamma_{k+1}}{\gamma_k + \delta_{k+1}\mu}.$$

Then, the next lemma will be used to prove that $\mathbb{E}[F(x_k)] \leq \mathbb{E}[d_k^*] + \xi_k$, where ξ_k is a noise term, such that $\mathbb{E}[F(x_k)] - F^* \leq \Gamma_k(d_0(x^*) - F^*) + \xi_k$, according to (8).

Lemma 2 (**Key lemma for acceleration**). *Consider the same sequences as in Lemma 1. Then, for all $k \geq 1$,*

$$\mathbb{E}[F(x_k)] \leq \mathbb{E}\left[l_k(y_{k-1})\right] + \left(\frac{L\eta_k^2}{2} - \eta_k\right)\mathbb{E}\left[\|\tilde{g}_k\|^2\right] + \eta_k\sigma_k^2,$$

with $\sigma_k^2 = \mathbb{E}[\|\nabla f(y_{k-1}) - g_k\|^2]$ and $\tilde{g}_k = g_k + \psi'(x_k)$.

Finally, we obtain the following convergence result.

Theorem 3 (**Convergence of accelerated SGD**). *Under the assumptions of Lemma 1, we have for all $k \geq 1$,*

$$\mathbb{E}\left[F(x_k) - F^* + \frac{\gamma_k}{2}\|v_k - x^*\|^2\right] \leq \Gamma_k\left(F(x_0) - F^* + \frac{\gamma_0}{2}\|x_0 - x^*\|^2 + \sum_{t=1}^{k}\frac{\eta_t\sigma_t^2}{\Gamma_t}\right),$$

where, as before, $\Gamma_t = \prod_{i=1}^t(1-\delta_i)$.

We now specialize the theorem to various practical cases.

Corollary 5 (**Prox accelerated SGD with constant η**). *Assume that g_k has constant variance $\sigma_k = \sigma$, and choose $\gamma_0 = \mu$ and $\eta_k = 1/L$ with Algorithm (B). Then,*

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k\left(F(x_0) - F^* + \frac{\mu}{2}\|x_0 - x^*\|^2\right) + \frac{\sigma^2}{\sqrt{\mu L}}.$$

We now show that with decreasing step sizes, we obtain an algorithm with optimal complexity similar to (Ghadimi & Lan, 2013; Cohen et al., 2018; Aybat et al., 2019), though we use a two-stages algorithm only.

Corollary 6 (**Prox accelerated SGD with decreasing η_k**). *Target an accuracy ε smaller than $2\sigma^2/\sqrt{\mu L}$. First, use a constant step-size $\eta_k = 1/L$ with $\gamma_0 = \mu$ within Algorithm (B) until $\mathbb{E}[F(x_k) - F^*] \leq 2\sigma^2/\sqrt{\mu L}$. Then, we restart the optimization procedure with decreasing step-sizes $\eta_k = \min\left(\frac{1}{L}, \frac{4}{\mu(k+2)^2}\right)$. The number of gradient evaluations to achieve $\mathbb{E}[F(x_k) - F^*] \leq \varepsilon$ is upper bounded by*

$$O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{F(x_0) - F^*}{\varepsilon}\right)\right) + O\left(\frac{\sigma^2}{\mu\varepsilon}\right).$$

4.2. Accelerated Algorithm with Variance Reduction

Next, we show how to build accelerated algorithms with the random-SVRG gradient estimator. First, we control the variance of the estimator in a similar manner to Katyusha (Allen-Zhu, 2017), as stated in the next proposition. Note that the estimator here does not require storing the seed of the random perturbations, unlike in the previous section, and does not rely on an averaging procedure (hence preserving the potential sparsity of the solution when ψ is sparsity-inducing).

Algorithm 2 Accelerated and robust random-SVRG

1: **Input:** x_0 in $\mathbb{R}^p$ (initial point); K (number of iterations); $(\eta_k)_{k\geq 0}$ (step sizes); $\gamma_0 \geq \mu$;
2: **Initialization:** $\tilde{x}_0 = v_0 = x_0$; $\bar{z}_0 = \tilde{\nabla} f(x_0)$;
3: **for** $k = 1, \ldots, K$ **do**
4: Find (δ_k, γ_k) such that

$$\gamma_k = (1-\delta_k)\gamma_{k-1} + \delta_k \mu \quad \text{and} \quad \delta_k = \sqrt{\frac{5\eta_k\gamma_k}{3n}};$$

5: Choose

$$y_{k-1} = \theta_k v_{k-1} + (1-\theta_k)\tilde{x}_{k-1} \text{ with } \theta_k = \frac{3n\delta_k - 5\mu\eta_k}{3-5\mu\eta_k};$$

6: Sample $i_k \sim Q = \{q_1, \ldots, q_n\}$;
7: Compute the gradient estimator with perturbations:

$$g_k = \frac{1}{q_{i_k} n}\left(\tilde{\nabla} f_{i_k}(y_{k-1}) - \tilde{\nabla} f_{i_k}(\tilde{x}_{k-1})\right) + \bar{z}_{k-1};$$

8: Obtain the new iterate

$$x_k \leftarrow \text{Prox}_{\eta_k\psi}\left[y_{k-1} - \eta_k g_k\right];$$

9: Find the minimizer v_k of the estimate sequence d_k:

$$v_k = \left(1 - \frac{\mu\delta_k}{\gamma_k}\right) v_{k-1} + \frac{\mu\delta_k}{\gamma_k} y_{k-1} + \frac{\delta_k}{\gamma_k\eta_k}(x_k - y_{k-1});$$

10: With probability $1/n$, update the anchor point

$$\tilde{x}_k = x_k \quad \text{and} \quad \bar{z}_k = \tilde{\nabla} f(\tilde{x}_k);$$

11: Otherwise, with probability $1-1/n$, keep the anchor point unchanged $\tilde{x}_k = \tilde{x}_{k-1}$ and $\bar{z}_k = \bar{z}_{k-1}$;
12: **end for**
13: **Output:** $\tilde{x}_k$.

Proposition 4 (Variance reduction for random-SVRG). *Consider problem (1) when f is a finite sum of functions $f = \frac{1}{n}\sum_{i=1}^n f_i$ where each f_i is L_i-smooth with $L_i \geq \mu$. Then, the variance of g_k defined in Algorithm 2 satisfies*

$$\sigma_k^2 \leq 2L_Q\left[f(\tilde{x}_{k-1}) - f(y_{k-1}) - g_k^\top(\tilde{x}_{k-1} - y_{k-1})\right] + 3\rho_Q\tilde{\sigma}^2.$$

Then, we extend Lemma 2 to the variance-reduction setting.

Lemma 3 (Key for accelerated variance-reduction). *Consider the iterates provided by Algorithm 2 and call $a_k = 2L_Q\eta_k$ and $\tilde{g}_k = g_k + \psi'(x_k)$. Then,*

$$\mathbb{E}[F(x_k)] \leq \mathbb{E}\left[a_k F(\tilde{x}_{k-1}) + (1-a_k) l_k(y_{k-1})\right]$$
$$+\mathbb{E}\left[a_k \tilde{g}_k^\top(y_{k-1} - \tilde{x}_{k-1}) + \left(\frac{L\eta_k^2}{2} - \eta_k\right)\|\tilde{g}_k\|^2\right] + 3\rho_Q\eta_k\tilde{\sigma}^2.$$

Then, we may now state our main convergence result.

Theorem 4 (Convergence of the accelerated SVRG). *Consider the iterates provided by Algorithm 2 and assume that $\eta_k \leq \min\left(\frac{1}{3L_Q}, \frac{1}{15\gamma_k n}\right)$ for all $k \geq 1$. Then,*

$$\mathbb{E}\left[F(x_k) - F^* + \frac{\gamma_k}{2}\|v_k - x^*\|^2\right]$$
$$\leq \Gamma_k\left(F(x_0) - F^* + \frac{\gamma_0}{2}\|x_0 - x^*\|^2 + \frac{3\rho_Q\tilde{\sigma}^2}{n}\sum_{t=1}^k \frac{\eta_t}{\Gamma_t}\right).$$

We may now derive convergence rates of our accelerated SVRG algorithm under various settings.

Corollary 7 (Accelerated prox SVRG with constant η). *With $\eta_k = \min\left(\frac{1}{3L_Q}, \frac{1}{15\mu n}\right)$ and $\gamma_0 = \mu$, the iterates produced by Algorithm 2 satisfy*

- *if $\frac{1}{3L_Q} \leq \frac{1}{15\mu n}$,*

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \left(1 - \sqrt{\frac{5\mu}{9L_Q n}}\right)^k T_0 + \frac{3\rho_Q\tilde{\sigma}^2}{\sqrt{5\mu L_Q n}};$$

- *otherwise,*

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \left(1 - \frac{1}{3n}\right)^k T_0 + \frac{3\rho_Q\tilde{\sigma}^2}{5\mu n},$$

with $T_0 = F(x_0) - F^* + \frac{\mu}{2}\|x_0 - x^*\|^2$.

The corollary uses $\Gamma_k\sum_{t=1}^k \eta/\Gamma_t \leq \eta/\delta = \sqrt{3n\eta/5\mu}$ and thus the algorithm converges linearly to an area of radius $3\rho_Q\tilde{\sigma}^2\sqrt{3\eta/5\mu n} = O\left(\rho_Q\tilde{\sigma}^2\min\left(\frac{1}{\sqrt{n\mu L_Q}}, \frac{1}{\mu n}\right)\right)$, where as before, $\rho_Q = 1$ if Q is uniform. When $\tilde{\sigma}^2 = 0$, the algorithm achieves the optimal complexity for finite sums (Arjevani & Shamir, 2016). Interestingly, we see that here non-uniform sampling may hurt the convergence guarantees in some situations. Whenever $5\mu n > \max_i L_i$, the optimal sampling strategy is indeed the uniform one. Next, we show how to obtain a converging algorithm.

Corollary 8 (Accelerated prox SVRG - decreasing η_k). *Target an accuracy ε smaller than $B = 3\rho_Q\tilde{\sigma}^2\sqrt{\eta/\mu}$ with the same step size η as in the previous corollary. First, use such a constant step-size strategy $\eta_k = \eta$ with $\gamma_0 = \mu$ within Algorithm 2, until $\mathbb{E}[F(x_k) - F^*] \leq B$. Then, restart the optimization procedure with decreasing step-sizes $\eta_k = \min\left(\eta, \frac{12n}{5\mu(k+2)^2}\right)$. The number of gradient evaluations to achieve $\mathbb{E}[F(x_k) - F^*] \leq \varepsilon$ is upper bounded by*

$$O\left(\left(n + \sqrt{\frac{nL_Q}{\mu}}\right)\log\left(\frac{F(x_0) - F^*}{\varepsilon}\right)\right) + O\left(\frac{\rho_Q\sigma^2}{\mu\varepsilon}\right).$$

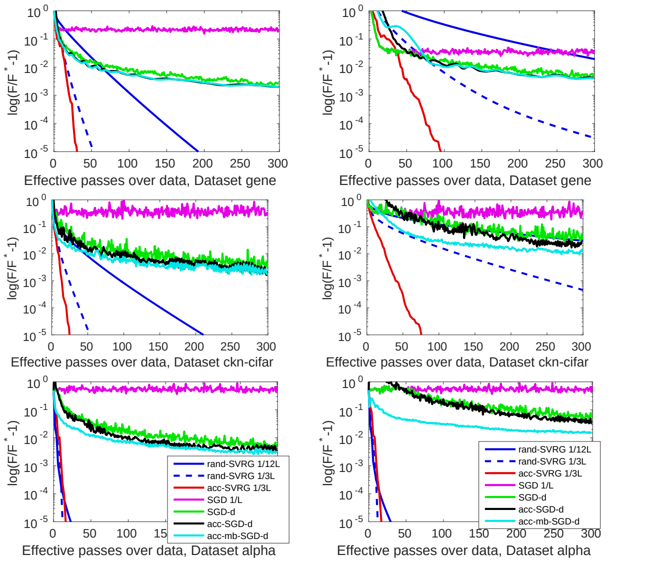

Figure 1. Objective function value on a logarithmic scale with $\lambda = 1/10n$ (left) and $\lambda = 1/100n$ (right), with no DropOut.

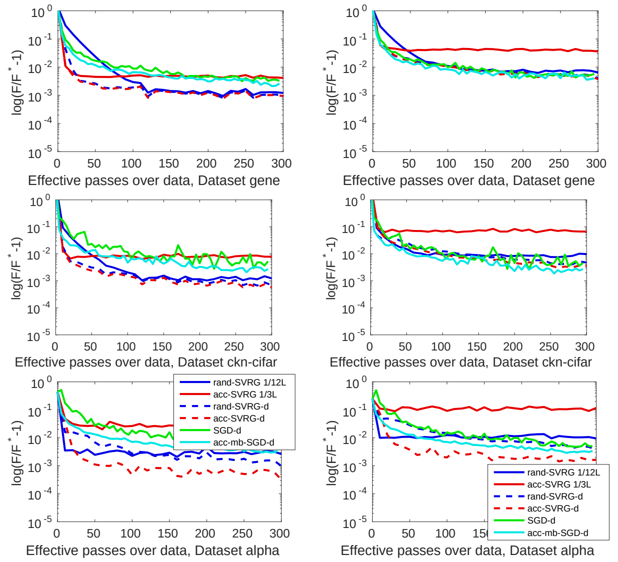

Figure 2. Objective function value on a logarithmic scale with $\lambda = 1/10n$, with DropOut $\delta = 0.01$ (left) and $\delta = 0.1$ (right).

5. Experiments

Following Bietti & Mairal (2017); Zheng & Kwok (2018) we consider logistic regression with DropOut (Srivastava et al., 2014), which consists of randomly setting to zero each vector entry with probability δ, leading to the problem

$$\min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\rho \left[\log(1 + e^{-b_i(\rho \circ a_i)^\top x}) \right] + \frac{\lambda}{2}\|x\|^2, \quad (16)$$

where ρ is a vector in $\{0,1\}^p$ with i.i.d. Bernoulli entries, $\circ$ denotes the elementwise multiplication between two vectors, the a_i's are vectors in $\mathbb{R}^p$ and b_i are labels in $\{-1,+1\}$. Since we normalize the vectors a_i, the corresponding functions f_i are L-smooth with $L = 0.25$. We consider two DropOut regimes, with δ in $\{0.01, 0.1\}$, representing small and medium perturbations. The parameter λ acts as a lower bound on μ and we consider $\lambda = 1/10n$, which is of the order of the smallest value that one would try when doing parameter search. We use three data sets alpha, ckn-cifar, and gene from different nature, which are presented in the appendix, along with other experimental details.

We consider various methods such as SGD, rand-SVRG, acc-SGD (accelerated SGD), and acc-SVRG (accelerated SVRG). We use them always with their theoretical step size, except rand-SVRG, which we evaluate with $\eta = 1/3L$ in order to obtain a fair comparison with acc-SVRG. When using the decreasing step size strategy, we add the suffix -d to the method's name, and we consider also a minibatch variant of acc-SGD, denoted by acc-mb-SGD with minibatch size $b = \sqrt{L/\mu}$. We also use the initial step size $1/3L$ for rand-SVRG-d since it performs better in practice. The methods do not use averaging, since it empirically slows down convergence when used from the first iteration; knowing when to start averaging is indeed not easy and requires heuristics which we do not evaluate here.

Experiments without perturbation (Figure 1). In such a regime, we obtain the following conclusions:
- Acceleration for SVRG is effective on gene and ckn-cifar except on alpha, where all SVRG-like methods perform already well. This may be due to hidden strong convexity leading to a regime where the complexity is $O(n \log(1/\varepsilon))$, which is independent of the condition number.
- Acceleration is more effective when the problem is badly conditioned—that is, when $\lambda = 1/100n$.
- acc-mb-SGD-d performs best among SGD methods and is competitive with rand-SVRG in the low precision regime.

Experiments with perturbations (Figure 2). As predicted by theory, approaches with constant step size do not converge. Therefore, we focus on methods with decreasing step sizes. The conclusions are the following:
- acc-mb-SGD-d with minibatch performs best among SGD approaches and could further benefit from parallelization.
- Acceleration for SVRG is less effective when DropOut is used; the gains are significant on the data set alpha, and the performance is similar as rand-SVRG on the two other data sets. Not reported here, acceleration is also more effective with poorly conditioned problems, when $\lambda = 1/100n$.
- acc-rand-SVRG-d performs better than SGD approaches in the low perturbation regime $\delta = 0.01$ and only on the alpha data set when $\delta = 0.1$. Otherwise, the methods perform similarly, making acc-rand-SVRG-d safe to use.

Acknowledgements

This work was supported by the ERC grant number 714381 (SOLARIS project). The authors would like to thank Anatoli Juditksy for interesting comments, and the anonymous reviewers who helped improving the manuscript.

References

Agarwal, A., Wainwright, M. J., Bartlett, P. L., and Ravikumar, P. K. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of Symposium on Theory of Computing (STOC)*, 2017.

Arjevani, Y. and Shamir, O. Dimension-free iteration complexity of finite sum optimization problems. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Aybat, N. S., Fallah, A., Gurbuzbalaban, M., and Ozdaglar, A. A universally optimal multistage accelerated stochastic gradient method. *preprint arXiv:1901.08022*, 2019.

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

Bietti, A. and Mairal, J. Stochastic optimization with variance reduction for infinite datasets with finite-sum structure. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

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

Cohen, M. B., Diakonikolas, J., and Orecchia, L. On acceleration with noise-corrupted gradients. In *Proceedings of the International Conferences on Machine Learning (ICML)*, 2018.

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 (NIPS)*, 2014.

Devolder, O. Stochastic first order methods in smooth convex optimization. CORE Discussion Papers 2011070, Universit catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2011.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization II: Shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

Hiriart-Urruty, J.-B. and Lemaréchal, C. *Convex analysis and minimization algorithms. II.* Springer, 1996.

Hofmann, T., Lucchi, A., Lacoste-Julien, S., and McWilliams, B. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems (NIPS)*, 2015a.

Hofmann, T., Lucchi, A., and McWilliams, B. Neighborhood watch: Stochastic gradient descent with neighbors. *CoRR*, abs/1506.03662, 2015b.

Hu, C., Pan, W., and Kwok, J. T. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems (NIPS)*. 2009.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

Kovalev, D., Horvath, S., and Richtarik, P. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. *preprint arXiv:1901.08689*, 2019.

Kulunchakov, A. and Mairal, J. Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise. *preprint arXiv:1901.08788*, 2019.

Lan, G. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

Lan, G. and Zhou, Y. An optimal randomized incremental gradient method. *Mathematical Programming*, 171(1–2): 167–215, 2018a.

Lan, G. and Zhou, Y. Random gradient extrapolation for distributed and stochastic optimization. *SIAM Journal on Optimization*, 28(4):2753–2782, 2018b.

Lei, L., Ju, C., Chen, J., and Jordan, M. I. Non-convex finite-sum optimization via SCSG methods. In *Advances in Neural Information Processing Systems (NIPS)*. 2017.

Lin, Q., Chen, X., and Peña, J. A sparsity preserving stochastic gradient methods for sparse regression. *Computational Optimization and Applications*, 58(2):455–482, 2014.

Lu, Z. and Xiao, L. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, 152(1):615–642, 2015.

Mairal, J. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

Mairal, J. End-to-end kernel learning with supervised convolutional kernel networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Meinshausen, N. and Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Moreau, J.-J. Fonctions convexes duales et points proximaux dans un espace hilbertien. *CR Acad. Sci. Paris Sér. A Math*, 255:2897–2899, 1962.

Moreau, J.-J. Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93(2):273–299, 1965.

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

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

Nesterov, Y. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the International Conferences on Machine Learning (ICML)*, 2017.

Nguyen, L. M., Nguyen, P. H., van Dijk, M., Richtárik, P., Scheinberg, K., and Takáč, M. SGD and Hogwild! convergence without the bounded gradients assumption. In *Proceedings of the International Conferences on Machine Learning (ICML)*, 2018.

Paquette, C., Lin, H., Drusvyatskiy, D., Mairal, J., and Harchaoui, Z. Catalyst acceleration for gradient-based non-convex optimization. *preprint arXiv:1703.10993*, 2018.

Schmidt, M., Babanezhad, R., Ahmed, M., Defazio, A., Clifton, A., and Sarkar, A. Non-uniform stochastic average gradient method for training conditional random fields. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Shalev-Shwartz, S. and Zhang, T. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 155(1):105–145, 2016.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 15(1):1929–1958, 2014.

Wainwright, M. J., Jordan, M. I., and Duchi, J. C. Privacy aware learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

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

Zheng, S. and Kwok, J. T. Lightweight stochastic optimization for minimizing finite sums with infinite data. In *Proceedings of the International Conferences on Machine Learning (ICML)*, 2018.

Zheng, S., Song, Y., Leung, T., and Goodfellow, I. Improving the robustness of deep neural networks via stability training. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Zhou, K. Direct acceleration of SAGA using sampled negative momentum. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Zhou, K., Shang, F., and Cheng, J. A simple stochastic variance reduced algorithm with fast convergence rates. In *Proceedings of the International Conferences on Machine Learning (ICML)*, 2018.

Supplementary Material of
"Estimate Sequences for Variance-Reduced Stochastic Composite Optimization"

A. Making SAGA Robust to Stochastic Perturbations

Algorithm 3 Iteration (A) with SAGA estimator

1: **Input:** x_0 in $\mathbb{R}^p$ (initial point); K (number of iterations); $(\eta_k)_{k\geq 0}$ (step sizes); $\beta \in [0, \mu]$; if averaging, $\gamma_0 \geq \mu$.
2: **Initialization:** $z_0^i = \tilde{\nabla} f_i(x_0) - \beta x_0$ for all $i = 1, \ldots, n$ and $\bar{z}_0 = \frac{1}{n}\sum_{i=1}^n z_0^i$.
3: **for** $k = 1, \ldots, K$ **do**
4: Sample i_k according to the distribution $Q = \{q_1, \ldots, q_n\}$;
5: Compute the gradient estimator, possibly corrupted by random perturbations:

$$g_k = \frac{1}{q_{i_k} n}\left(\tilde{\nabla} f_{i_k}(x_{k-1}) - \beta x_{k-1} - z_{k-1}^{i_k}\right) + \bar{z}_{k-1} + \beta x_{k-1};$$

6: Obtain the new iterate

$$x_k \leftarrow \mathrm{Prox}_{\eta_k \psi}\left[x_{k-1} - \eta_k g_k\right];$$

7: Draw j_k from the uniform distribution in $\{1, \ldots, n\}$;
8: Update the auxiliary variables

$$z_k^{j_k} = \tilde{\nabla} f_{j_k}(x_k) - \beta x_k \quad \text{and} \quad z_k^j = z_{k-1}^j \quad \text{for all} \quad j \neq j_k;$$

9: Update the average variable $\bar{z}_k = \bar{z}_{k-1} + \frac{1}{n}(z_k^{j_k} - z_{k-1}^{j_k})$.
10: **Optional**: Use the same averaging strategy as in Algorithm 1.
11: **end for**
12: **Output:** x_k or $\hat{x}_k$ (if averaging).

B. Details about the Experimental Setup

We consider three datasets with various number of points n and dimension p, coming from different scientific fields:

- alpha is from the Pascal Large Scale Learning Challenge website[1] and contains $n = 250\,000$ with $p = 500$.
- gene consists of gene expression data and the binary labels b_i characterize two different types of breast cancer. This is a small dataset with $n = 295$ and $p = 8\,141$.
- ckn-cifar is an image classification task where each image from the CIFAR-10 dataset[2] is represented by using a two-layer unsupervised convolutional neural network (Mairal, 2016). Since CIFAR-10 originally contains 10 different classes, we consider the binary classification task consisting of predicting the class 1 vs. other classes. The dataset contains $n = 50\,000$ images and the dimension of the representation is $p = 9\,216$.

For simplicity, we normalize the features of all datasets and thus we use a uniform sampling strategy Q in all algorithms. Then, we consider several methods with their theoretical step sizes, described in Table 1. Note that we also evaluate the strategy random-SVRG with step size $1/3L$, even though our analysis requires $1/12L$, in order to get a fair comparison with the accelerated SVRG method. In all figures, we consider that n iterations of SVRG count as 2 effective passes over the data since it appears empirically to be a good proxy of the computational time. Indeed, (i) if one is allowed to store all variables z_i^k, then n iterations indeed correspond to two passes over the data; (ii) the gradients $\tilde{\nabla} f_i(x_{k-1}) - \tilde{\nabla} f_i(\tilde{x}_{k-1})$ access the same training point which reduces the data access overhead; (iii) computing the full gradient $\bar{z}_k$ can be done in practice in a much more efficient manner than computing individually the n gradients $\tilde{\nabla} f_i(x_k)$, either through parallelization

[1] http://largescale.ml.tu-berlin.de/
[2] https://www.cs.toronto.edu/~kriz/cifar.html

or by using more efficient routines (*e.g.*, BLAS2). Each experiment is conducted five times and we always report the average of the five experiments in each figure.

To evaluate the quality of a solution, when $\tilde{\sigma}^2 = 0$, we can check that the value F^* we consider is optimal by computing a duality gap using Fenchel duality. In the stochastic case when $\tilde{\sigma}^2 \neq 0$, we evaluate the loss function every 5 data passes and we estimate the expectation (16) by drawing 5 random perturbations per data point, resulting in $5n$ samples. The optimal value F^* is estimated by letting the methods run for 1000 epochs and selecting the best point found as a proxy of F^*.

Algorithm	step size η_k	Theory	Complexity $O(.)$	Bias $O(.)$
SGD	$\frac{1}{L}$	Cor. 1	$\frac{L}{\mu}\log\left(\frac{C_0}{\varepsilon}\right)$	$\frac{\sigma^2}{L}$
SGD-d	$\min\left(\frac{1}{L}, \frac{2}{\mu(k+2)}\right)$	Cor. 2	$\frac{L}{\mu}\log\left(\frac{C_0}{\varepsilon}\right) + \frac{\sigma^2}{\mu\varepsilon}$	0
acc-SGD	$\frac{1}{L}$	Cor. 5	$\sqrt{\frac{L}{\mu}}\log\left(\frac{C_0}{\varepsilon}\right)$	$\frac{\sigma^2}{\sqrt{\mu L}}$
acc-SGD-d	$\min\left(\frac{1}{L}, \frac{4}{\mu(k+2)^2}\right)$	Cor. 6	$\sqrt{\frac{L}{\mu}}\log\left(\frac{C_0}{\varepsilon}\right) + \frac{\sigma^2}{\mu\varepsilon}$	0
acc-mb-SGD-d	$\min\left(\frac{1}{L}, \frac{4}{\mu(k+2)^2}\right)$	Cor. 6	$\frac{L}{\mu}\log\left(\frac{C_0}{\varepsilon}\right) + \frac{\sigma^2}{\mu\varepsilon}$	0
rand-SVRG	$\frac{1}{12L}$	Cor. 3	$\left(n + \frac{L}{\mu}\right)\log\left(\frac{C_0}{\varepsilon}\right)$	$\frac{\tilde{\sigma}^2}{L}$
rand-SVRG-d	$\min\left(\frac{1}{12L_Q}, \frac{1}{5\mu n}, \frac{2}{\mu(k+2)}\right)$	Cor. 4	$\left(n + \frac{L}{\mu}\right)\log\left(\frac{C_0}{\varepsilon}\right) + \frac{\tilde{\sigma}^2}{\mu\varepsilon}$	0
acc-SVRG	$\min\left(\frac{1}{3L_Q}, \frac{1}{15\mu n}\right)$	Cor. 7	$\left(n + \sqrt{\frac{nL}{\mu}}\right)\log\left(\frac{C_0}{\varepsilon}\right)$	$\frac{\tilde{\sigma}^2}{\sqrt{n\mu L}+n\mu}$
acc-SVRG-d	$\min\left(\frac{1}{3L_Q}, \frac{1}{15\mu n}, \frac{12n}{5\mu(k+2)^2}\right)$	Cor. 8	$\left(n + \sqrt{\frac{nL}{\mu}}\right)\log\left(\frac{C_0}{\varepsilon}\right) + \frac{\tilde{\sigma}^2}{\mu\varepsilon}$	0

Table 1. List of algorithms used in the experiments, along with the step size used and the pointer to the corresponding convergence guarantees, with $C_0 = F(x_0) - F^*$. In the experiments, we also use the method rand-SVRG with step size $\eta = 1/3L$. The approach acc-mb-SGD-d uses minibatches of size $\lceil\sqrt{L/\mu}\rceil$ and could thus easily be parallelized. Note that we potentially have $\tilde{\sigma} \ll \sigma$.

C. Useful Mathematical Results

C.1. Simple Results about Convexity and Smoothness

The next three lemmas are classical upper and lower bounds for smooth or strongly convex functions (Nesterov, 2004).

Lemma C.1 (Quadratic upper bound for L-smooth functions).
Let $f : \mathbb{R}^p \to \mathbb{R}$ be L-smooth. Then, for all x, x' in $\mathbb{R}^p$,

$$|f(x') - f(x) - \nabla f(x)^\top (x' - x)| \leq \frac{L}{2}\|x - x'\|^2.$$

Lemma C.2 (Lower bound for strongly convex functions).
Let $f : \mathbb{R}^p \to \mathbb{R}$ be a μ-strongly convex function. Let z be in $\partial f(x)$ for some x in $\mathbb{R}^p$. Then, the following inequality holds for all x' in $\mathbb{R}^p$:

$$f(x') \geq f(x) + z^\top (x' - x) + \frac{\mu}{2}\|x - x'\|^2.$$

Lemma C.3 (Second-order growth property).
Let $f : \mathbb{R}^p \to \mathbb{R}$ be a μ-strongly convex function and $\mathcal{X} \subseteq \mathbb{R}^p$ be a convex set. Let x^* be the minimizer of f on $\mathcal{X}$. Then, the following condition holds for all x in $\mathcal{X}$:

$$f(x) \geq f(x^*) + \frac{\mu}{2}\|x - x^*\|^2.$$

C.2. Useful Results to Select Step Sizes

In this section, we present basic mathematical results regarding the choice of step sizes. The proof of the first two lemmas is trivial by induction.

Lemma C.4 (Relation between $(\delta_k)_{k\geq 0}$ and $(\Gamma_k)_{k\geq 0}$)**.** *Consider the following scenarios for δ_k and $\Gamma_k = \prod_{t=1}^k (1 - \delta_t)$:*

- *$\delta_k = \delta$ (constant). Then $\Gamma_k = (1 - \delta)^k$.*

- $\delta_k = 2/(k+2)$. *Then,* $\Gamma_k = \frac{2}{(k+1)(k+2)}$.
- $\delta_k = \min(2/(k+2), \delta)$. *Then,*

$$\Gamma_k = \begin{cases} (1-\delta)^k & \text{if } k < k_0 \quad \text{with} \quad k_0 = \left\lceil \frac{2}{\delta} - 2 \right\rceil \\ \Gamma_{k_0-1}\frac{k_0(k_0+1)}{(k+1)(k+2)} & \text{otherwise.} \end{cases}$$

Lemma C.5 (Simple relation)**.** *Consider a sequence of weights $(\delta_k)_{k\geq 0}$ in $(0,1)$. Then,*

$$\sum_{t=1}^{k} \frac{\delta_t}{\Gamma_t} + 1 = \frac{1}{\Gamma_k} \qquad \text{where} \qquad \Gamma_t := \prod_{i=1}^{t}(1-\delta_i). \tag{17}$$

Lemma C.6 (Convergence rate of Γ_k)**.** *Consider the same quantities defined in the previous lemma and consider the sequence $\gamma_k = (1-\delta_k)\gamma_{k-1} + \delta_k \mu = \Gamma_k \gamma_0 + (1-\Gamma_k)\mu$ with $\gamma_0 \geq \mu$, and assume the relation $\delta_k = \gamma_k \eta$. Then, for all $k \geq 0$,*

$$\Gamma_k \leq \min\left((1-\mu\eta)^k, \frac{1}{1+\gamma_0 \eta k} \right). \tag{18}$$

Besides,

- *when $\gamma_0 = \mu$, then $\Gamma_k = (1-\mu\eta)^k$.*
- *when $\mu = 0$, $\Gamma_k = \frac{1}{1+\gamma_0\eta k}$.*

Proof. First, we have for all k, $\gamma_k \geq \mu$ such that $\delta_k \geq \eta\mu$, which leads then to $\Gamma_k \leq (1-\eta\mu)^k$. Besides, $\gamma_k \geq \Gamma_k \gamma_0$ and thus $\Gamma_k = (1-\delta_k)\Gamma_{k-1} \leq (1-\Gamma_k\gamma_0\eta)\,\Gamma_{k-1}$. Then, $\frac{1}{\Gamma_k}(1-\Gamma_k\gamma_0\eta) \geq \frac{1}{\Gamma_{k-1}}$, and

$$\frac{1}{\Gamma_k} \geq \frac{1}{\Gamma_{k-1}} + \gamma_0\eta \geq 1 + \gamma_0 \eta k,$$

which is sufficient to obtain (18). Then, the fact that $\gamma_0 = \mu$ leads to $\Gamma_k = (1-\mu\eta)^k$ is trivial, and the fact that $\mu = 0$ yields $\Gamma_k = \frac{1}{1+\gamma_0\eta k}$ can be shown by induction. Indeed, the relation is true for Γ_0 and then, assuming the relation is true for $k-1$, we have for $k \geq 1$,

$$\Gamma_k = (1-\delta_k)\Gamma_{k-1} = (1-\eta\gamma_k)\Gamma_{k-1} = (1-\eta\gamma_0\Gamma_k)\Gamma_{k-1} \geq (1-\eta\gamma_0\Gamma_k)\frac{1}{1+\gamma_0\eta(k-1)},$$

which leads to $\Gamma_k = \frac{1}{1+\gamma_0\eta k}$. □

Lemma C.7 (Accelerated convergence rate of Γ_k)**.** *Consider the same quantities defined in Lemma C.5 and consider the sequence $\gamma_k = (1-\delta_k)\gamma_{k-1} + \delta_k\mu = \Gamma_k\gamma_0 + (1-\Gamma_k)\mu$ with $\gamma_0 \geq \mu$, and assume the relation $\delta_k = \sqrt{\gamma_k \eta}$. Then, for all $k \geq 0$,*

$$\Gamma_k \leq \min\left((1-\sqrt{\mu\eta})^k, \frac{4}{(2+\sqrt{\gamma_0\eta}k)^2} \right).$$

Besides, when $\gamma_0 = \mu$, then $\Gamma_k = (1-\sqrt{\mu\eta})^k$.

Proof. see Lemma 2.2.4 of (Nesterov, 2004). □

C.3. Averaging Strategy

Next, we show a generic convergence result and an appropriate averaging strategy given a recursive relation between quantities acting as Lyapunov function.

Lemma C.8 (Averaging strategy)**.** *Assume that an algorithm generates a sequence $(x_k)_{k\geq 0}$ for minimizing a convex function F, and that there exist sequences $(T_k)_{k\geq 0}$, $(\delta_k)_{k\geq 1}$ in $(0,1)$, $(\beta_k)_{k\geq 1}$ and a scalar $\alpha > 0$ such that for all $k \geq 1$,*

$$\frac{\delta_k}{\alpha}\mathbb{E}[F(x_k) - F^*] + T_k \leq (1-\delta_k)T_{k-1} + \beta_k, \tag{19}$$

where the expectation is taken with respect to any random parameter used by the algorithm. Then, we consider two cases:

No averaging.

$$\mathbb{E}[F(x_k) - F^*] + \frac{\alpha}{\delta_k} T_k \leq \frac{\alpha \Gamma_k}{\delta_k} \left(T_0 + \sum_{t=1}^{k} \frac{\beta_t}{\Gamma_t} \right) \quad \textit{where} \quad \Gamma_k := \prod_{t=1}^{k} (1 - \delta_t). \tag{20}$$

Averaging. *By defining the averaging sequence* $(\hat{x}_k)_{k \geq 0}$,

$$\hat{x}_k = \Gamma_k \left(x_0 + \sum_{t=1}^{k} \frac{\delta_t}{\Gamma_t} x_t \right) = (1 - \delta_k)\hat{x}_{k-1} + \delta_k x_k \quad \textit{(for } k \geq 1\textit{)},$$

then,

$$\mathbb{E}[F(\hat{x}_k) - F^*] + \alpha T_k \leq \Gamma_k \left(\alpha T_0 + \mathbb{E}[F(x_0) - F^*] + \alpha \sum_{t=1}^{k} \frac{\beta_t}{\Gamma_t} \right). \tag{21}$$

Proof. Given that $T_k \leq (1 - \delta_k) T_{k-1} + \beta_k$, we obtain (20) by simply unrolling the recursion. To analyze the effect of the averaging strategies, divide now (19) by Γ_k:

$$\frac{\delta_k}{\alpha \Gamma_k} \mathbb{E}[F(x_k) - F^*] + \frac{T_k}{\Gamma_k} \leq \frac{T_{k-1}}{\Gamma_{k-1}} + \frac{\beta_k}{\Gamma_k}.$$

Sum from $t = 1$ to k and notice that we have a telescopic sum:

$$\frac{1}{\alpha} \sum_{t=1}^{k} \frac{\delta_t}{\Gamma_t} \mathbb{E}[F(x_t) - F^*] + \frac{T_k}{\Gamma_k} \leq T_0 + \sum_{t=1}^{k} \frac{\beta_t}{\Gamma_t}.$$

Then, add $(1/\alpha)\mathbb{E}[F(x_0) - F^*]$ on both sides and multiply by $\alpha \Gamma_k$:

$$\sum_{t=1}^{k} \frac{\delta_t \Gamma_k}{\Gamma_t} \mathbb{E}[F(x_t) - F^*] + \Gamma_k \mathbb{E}[F(x_0) - F^*] + \alpha T_k \leq \Gamma_k \left(\alpha T_0 + \mathbb{E}[F(x_0) - F^*] + \alpha \sum_{t=1}^{k} \frac{\beta_t}{\Gamma_t} \right).$$

By exploiting the relation (17), we may then use Jensen's inequality and we obtain (21). □

D. Proofs of the Main Results

D.1. Proof of Proposition 1

Proof.

$$\begin{aligned}
d_k^* = d_k(x_k) &= (1 - \delta_k) d_{k-1}(x_k) + \delta_k \left(f(x_{k-1}) + g_k^\top (x_k - x_{k-1}) + \frac{\mu}{2} \|x_k - x_{k-1}\|^2 + \psi(x_k) \right) \\
&\geq (1 - \delta_k) d_{k-1}^* + \frac{\gamma_k}{2} \|x_k - x_{k-1}\|^2 + \delta_k \left(f(x_{k-1}) + g_k^\top (x_k - x_{k-1}) + \psi(x_k) \right) \\
&\geq (1 - \delta_k) d_{k-1}^* + \delta_k \left(f(x_{k-1}) + g_k^\top (x_k - x_{k-1}) + \frac{L}{2} \|x_k - x_{k-1}\|^2 + \psi(x_k) \right) \\
&\geq (1 - \delta_k) d_{k-1}^* + \delta_k F(x_k) + \delta_k (g_k - \nabla f(x_{k-1}))^\top (x_k - x_{k-1}),
\end{aligned}$$

where the first inequality comes from Lemma C.3—it is in fact an equality when considering Algorithm (A)—and the second inequality simply uses the assumption $\eta_k \leq 1/L$, which yields $\delta_k = \gamma_k \eta_k \leq \gamma_k / L$. Finally, the last inequality uses a classical upper-bound for L-smooth functions presented in Lemma C.1. Then, after taking expectations,

$$\begin{aligned}
E[d_k^*] &\geq (1 - \delta_k) \mathbb{E}[d_{k-1}^*] + \delta_k \mathbb{E}[F(x_k)] + \delta_k \mathbb{E}[(g_k - \nabla f(x_{k-1}))^\top (x_k - x_{k-1})] \\
&= (1 - \delta_k) \mathbb{E}[d_{k-1}^*] + \delta_k \mathbb{E}[F(x_k)] + \delta_k \mathbb{E}[(g_k - \nabla f(x_{k-1}))^\top x_k] \\
&= (1 - \delta_k) \mathbb{E}[d_{k-1}^*] + \delta_k \mathbb{E}[F(x_k)] + \delta_k \mathbb{E}\left[(g_k - \nabla f(x_{k-1}))^\top (x_k - w_{k-1}) \right],
\end{aligned}$$

where we have defined the following quantity

$$w_{k-1} = \text{Prox}_{\eta_k \psi} \left[x_{k-1} - \eta_k \nabla f(x_{k-1}) \right].$$

In the previous relations, we have used twice the fact that $\mathbb{E}[(g_k - \nabla f(x_{k-1}))^\top y | \mathcal{F}_{k-1}] = 0$, for all y that is deterministic given x_{k-1} such as $y = x_{k-1}$ or $y = w_{k-1}$. We may now use the non-expansiveness property of the proximal operator (Moreau, 1965) to control the quantity $\|x_k - w_{k-1}\|$, which gives us

$$
\begin{aligned}
E[d_k^*] &\geq (1-\delta_k)\mathbb{E}[d_{k-1}^*] + \delta_k \mathbb{E}[F(x_k)] - \delta_k \mathbb{E}\left[\|g_k - \nabla f(x_{k-1})\| \|x_k - w_{k-1}\|\right] \\
&\geq (1-\delta_k)\mathbb{E}[d_{k-1}^*] + \delta_k \mathbb{E}[F(x_k)] - \delta_k \eta_k \mathbb{E}\left[\|g_k - \nabla f(x_{k-1})\|^2\right] \\
&= (1-\delta_k)\mathbb{E}[d_{k-1}^*] + \delta_k \mathbb{E}[F(x_k)] - \delta_k \eta_k \sigma_k^2.
\end{aligned}
$$

This relation can now be combined with (8) when $z = x^*$, and we obtain (11). $\square$

D.2. Proof of Corollary 2

Proof. Given the linear convergence rate (12), the number of iterations to guarantee $\mathbb{E}[F(\hat{x}_k) - F^*] \leq 2\sigma^2/L$ with the constant step-size strategy is upper bounded by

$$
O\left(\frac{L}{\mu}\log\left(\frac{F(x_0) - F^*}{\varepsilon}\right)\right).
$$

Then, after restarting the algorithm, we may apply Theorem 1 with $\mathbb{E}[F(x_0) - F^*] \leq 2\sigma^2/L$. With $\gamma_0 = \mu$, we have $\gamma_k = \mu$ for all $k \geq 0$, and the rate of Γ_k is given by Lemma C.4, which yields for $k \geq k_0 = \left\lceil \frac{2L}{\mu} - 2 \right\rceil$,

$$
\begin{aligned}
\mathbb{E}[F(\hat{x}_k) - F^*] &\leq \Gamma_k \left(\mathbb{E}\left[F(x_0) - F^* + \frac{\mu}{2}\|x_0 - x^*\|^2\right] + \sigma^2 \sum_{t=1}^{k} \frac{\delta_t \eta_t}{\Gamma_t} \right) \\
&\leq \Gamma_k \left(\frac{4\sigma^2}{L} + \frac{\sigma^2}{L} \sum_{t=1}^{k_0-1} \frac{\delta_t}{\Gamma_t} + \sigma^2 \sum_{t=k_0}^{k} \frac{2\delta_t}{\Gamma_t \mu (t+2)} \right) \\
&= \frac{k_0(k_0+1)}{(k+1)(k+2)} \left(\Gamma_{k_0-1}\frac{4\sigma^2}{L} + \frac{\sigma^2}{L}\Gamma_{k_0-1} \sum_{t=1}^{k_0-1} \frac{\delta_t}{\Gamma_t} \right) + \sigma^2 \sum_{t=k_0}^{k} \frac{2\delta_t \Gamma_k}{\Gamma_t \mu (t+2)} \\
&= \frac{k_0(k_0+1)}{(k+1)(k+2)} \left(\Gamma_{k_0-1}\frac{4\sigma^2}{L} + (1 - \Gamma_{k_0-1})\frac{\sigma^2}{L} \right) + \sigma^2 \sum_{t=k_0}^{k} \frac{2\delta_t \Gamma_k}{\Gamma_t \mu (t+2)} \\
&\leq \frac{k_0(k_0+1)}{(k+1)(k+2)} \frac{4\sigma^2}{L} + \sigma^2 \frac{1}{(k+1)(k+2)} \left(\sum_{t=k_0+1}^{k} \frac{4(t+1)(t+2)}{\mu(t+2)^2} \right) \\
&\leq \frac{k_0}{(k+1)(k+2)} \frac{8\sigma^2}{\mu} + \frac{4\sigma^2}{\mu(k+2)},
\end{aligned}
$$

where the second inequality uses the fact that $\frac{\mu}{2}\|x_0 - x^*\|^2 \leq F(x_0) - F^* \leq \frac{2\sigma^2}{L}$, and then we use Lemmas C.4 and C.5. The term on the right is of order $O(\sigma^2/\mu k)$ whereas the term on the left becomes of the same order or smaller whenever $k \geq k_0 = O(L/\mu)$. This leads to the desired iteration complexity. $\square$

D.3. Proof of Proposition 2

Proof. The proof borrows a large part of the analysis of Xiao & Zhang (2014) for controlling the variance of the gradient estimate in the SVRG algorithm. First, we note that all the gradient estimators we consider may be written as

$$
g_k = \frac{1}{q_{i_k} n}\left(\tilde{\nabla} f_{i_k}(x_{k-1}) - z_{k-1}^{i_k}\right) + \bar{z}_{k-1}.
$$

Then, we will write $\tilde{\nabla} f_{i_k}(x_{k-1}) = \nabla f_{i_k}(x_{k-1}) + \zeta_k$, where ζ_k is a zero-mean variable with variance $\tilde{\sigma}^2$ drawn at iteration k, and $z_k^i = u_k^i + \zeta_k^i$ for all k, i, where ζ_k^i has zero-mean with variance $\tilde{\sigma}^2$ and was drawn during the previous iterations. Then,

$$
\begin{aligned}
\sigma_k^2 &= \mathbb{E}\left\| \frac{1}{q_{i_k} n}(\tilde{\nabla} f_{i_k}(x_{k-1}) - z_{k-1}^{i_k}) + \bar{z}_{k-1} - \nabla f(x_{k-1}) \right\|^2 \\
&= \mathbb{E}\left\| \frac{1}{q_{i_k} n}(\nabla f_{i_k}(x_{k-1}) - z_{k-1}^{i_k}) + \bar{z}^{k-1} - \nabla f(x_{k-1}) \right\|^2 + \mathbb{E}\left[\frac{1}{(q_{i_k} n)^2} \|\zeta_k\|^2 \right] \\
&\leq \mathbb{E}\left\| \frac{1}{q_{i_k} n}(\nabla f_{i_k}(x_{k-1}) - z_{k-1}^{i_k}) + \bar{z}^{k-1} - \nabla f(x_{k-1}) \right\|^2 + \rho_Q \tilde{\sigma}^2 \\
&\leq \mathbb{E}\left\| \frac{1}{q_{i_k} n}(\nabla f_{i_k}(x_{k-1}) - z_{k-1}^{i_k}) \right\|^2 + \rho_Q \tilde{\sigma}^2 \\
&= \frac{1}{n}\sum_{i=1}^n \frac{1}{q_i n} \mathbb{E}\left[\|\nabla f_i(x_{k-1}) - z_{k-1}^i\|^2 \right] + \rho_Q \tilde{\sigma}^2 \\
&= \frac{1}{n}\sum_{i=1}^n \frac{1}{q_i n} \mathbb{E}\left[\|\nabla f_i(x_{k-1}) - u_*^i + u_*^i - z_{k-1}^i\|^2 \right] + \rho_Q \tilde{\sigma}^2 \quad \text{with} \quad u_i^* = \nabla f_i(x^*) \\
&\leq \frac{2}{n}\sum_{i=1}^n \frac{1}{q_i n} \mathbb{E}\left[\|\nabla f_i(x_{k-1}) - u_*^i\|^2 \right] + \frac{2}{n}\sum_{i=1}^n \frac{1}{q_i n} \mathbb{E}\left[\|z_{k-1}^i - u_*^i\|^2 \right] + \rho_Q \tilde{\sigma}^2 \\
&\leq \frac{2}{n}\sum_{i=1}^n \frac{1}{q_i n} \mathbb{E}\left[\|\nabla f_i(x_{k-1}) - \nabla f_i(x^*))\|^2 \right] + \frac{2}{n}\sum_{i=1}^n \frac{1}{q_i n} \mathbb{E}\left[\|u_{k-1}^i - u_*^i\|^2 \right] + 3\rho_Q \tilde{\sigma}^2 \\
&\leq \frac{4}{n}\sum_{i=1}^n \frac{L_i}{q_i n} \mathbb{E}\left[f_i(x_{k-1}) - f_i(x^*) - \nabla f_i(x^*)^\top (x_{k-1} - x^*) \right] + \frac{2}{n}\sum_{i=1}^n \frac{1}{q_i n} \mathbb{E}\left[\|u_{k-1}^i - u_*^i\|^2 \right] + 3\rho_Q \tilde{\sigma}^2 \\
&\leq 4L_Q \mathbb{E}\left[f(x_{k-1}) - f(x^*) - \nabla f(x^*)^\top (x_{k-1} - x^*) \right] + \frac{2}{n}\sum_{i=1}^n \frac{1}{q_i n} \mathbb{E}\left[\|u_{k-1}^i - u_*^i\|^2 \right] + 3\rho_Q \tilde{\sigma}^2,
\end{aligned}
$$

where the second inequality uses the relation $\mathbb{E}[\|X - \mathbb{E}[X]\|^2] \leq \mathbb{E}[\|X\|^2]$ for all random variable X, taking here expectation with respect to the index $i_k \sim Q$ and conditioning on $\mathcal{F}_{k-1}$; the third inequality uses the relation $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$; the fifth inequality uses Theorem 2.1.5 of (Nesterov, 2004).

Then, since x^* minimizes F, we have $0 \in \nabla f(x^*) + \partial\psi(x^*)$ and thus $-\nabla f(x^*)$ is a subgradient in $\partial\psi(x^*)$. By using as well the convexity inequality $\psi(x) \geq \psi(x^*) - \nabla f(x^*)^\top (x - x^*)$, we obtain

$$
f(x_{k-1}) - f(x^*) - \nabla f(x^*)^\top (x_{k-1} - x^*) \leq 2L_Q (F(x_{k-1}) - F^*).
$$

Finally, given the previous relations, we obtain (13). □

D.4. Proof of Proposition 3

Proof. To make the notation more compact, we call

$$
F_k = \mathbb{E}[F(x_k) - F^*], \qquad D_k = \mathbb{E}[d_k(x^*) - d_k^*] \qquad \text{and} \quad C_k = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \frac{1}{q_i n} \|u_k^i - u_*^i\|^2 \right].
$$

Then, according to Proposition 2, we have

$$
\sigma_k^2 \leq 4L_Q F_{k-1} + 2C_{k-1} + 3\rho_Q \tilde{\sigma}^2,
$$

and according to Proposition 1,

$$
\delta_k F_k + D_k \leq (1 - \delta_k) D_{k-1} + 4L_Q \eta_k \delta_k F_{k-1} + 2\eta_k \delta_k C_{k-1} + 3\rho_Q \eta_k \delta_k \tilde{\sigma}^2. \tag{22}
$$

Then, we note that both for the SVRG and SAGA, we have,

$$\mathbb{E}[\|u_k^i - u_*^i\|^2] = \left(1 - \frac{1}{n}\right)\mathbb{E}[\|u_{k-1}^i - u_*^i\|^2] + \frac{1}{n}\mathbb{E}\|\nabla f_i(x_k) - \nabla f_i(x^*)\|^2.$$

By taking a weighted average, this yields

$$\begin{aligned} C_k &\leq \left(1 - \frac{1}{n}\right) C_{k-1} + \frac{1}{n^2}\sum_{i=1}^n \frac{1}{q_i n}\mathbb{E}\left[\|\nabla f_i(x_k) - \nabla f_i(x^*)\|^2\right] \\ &\leq \left(1 - \frac{1}{n}\right) C_{k-1} + \frac{1}{n^2}\sum_{i=1}^n \frac{2L_i}{q_i n}\mathbb{E}\left[f_i(x_k) - f_i(x^*) - \nabla f_i(x^*)^\top (x_k - x^*)\right] \\ &\leq \left(1 - \frac{1}{n}\right) C_{k-1} + \frac{2L_Q F_k}{n}, \end{aligned}$$

where the second inequality comes from Theorem 2.1.5 of (Nesterov, 2004) and the last one uses similar arguments as in the proof of Proposition 2. Then, we add a quantity $\beta_k C_k$ on both sides of the relation (22) with some $\beta_k > 0$ that we will specify later:

$$\left(\delta_k - \beta_k \frac{2L_Q}{n}\right) F_k + D_k + \beta_k C_k \leq (1-\delta_k) D_{k-1} + \left(\beta_k \left(1 - \frac{1}{n}\right) + 2\eta_k\delta_k\right) C_{k-1} + 4L_Q\eta_k\delta_k F_{k-1} + 3\rho_Q\eta_k\delta_k\tilde{\sigma}^2,$$

and then choose $\frac{\beta_k}{n} = \frac{5}{2}\eta_k\delta_k$, which yields

$$\delta_k\left(1 - 5L_Q\eta_k\right) F_k + D_k + \beta_k C_k \leq (1-\delta_k) D_{k-1} + \beta_k \left(1 - \frac{1}{5n}\right) C_{k-1} + 4L_Q\eta_k\delta_k F_{k-1} + 3\rho_Q\eta_k\delta_k\tilde{\sigma}^2.$$

Remember that $\tau_k = \min\left(\delta_k, \frac{1}{5n}\right)$, notice that the sequences $(\beta_k)_{k\geq 0}$, $(\eta_k)_{k\geq 0}$ and $(\delta_k)_{k\geq 0}$ are non-increasing and note that $4 \leq 5(1 - \frac{1}{5n})$ for all $n \geq 1$. Then,

$$\delta_k\left(1 - 10L_Q\eta_k\right) F_k + \underbrace{5L_Q\eta_k\delta_k + D_k + \beta_k C_k}_{T_k} \leq (1-\tau_k)\left(D_{k-1} + \beta_{k-1}C_{k-1} + 5L_Q\eta_{k-1}\delta_{k-1}F_{k-1}\right) + 3\rho_Q\eta_k\delta_k\tilde{\sigma}^2,$$

which immediately yields (14) with the appropriate definition of T_k, and by noting that $(1 - 10L_Q\eta_k) \geq \frac{1}{6}$. □

D.5. Proof of Corollary 3

Proof. First, notice that (i) $T_k \geq d_k(x^*) - d_k^* \geq \frac{\mu}{2}\|x_k - x^*\|^2$, that (ii) $\delta_k = \eta_k\gamma_k = \frac{\mu}{12L_Q}$ and that $\mu\frac{\tau_k}{\delta_k} = \min\left(\mu, \frac{12L_Q}{5n}\right)$. Then, we apply Theorem 2 and obtain

$$\begin{aligned} \mathbb{E}\left[F(\hat{x}_k) - F^* + \alpha\|x_k - x^*\|^2\right] &\leq \Theta_k\left(F(x_0) - F^* + \frac{6\tau_k}{\delta_k}T_0 + \frac{18\rho_Q\tau_k\tilde{\sigma}^2}{\delta_k}\sum_{t=1}^k \frac{\eta_t\delta_t}{\Theta_t}\right) \\ &= \Theta_k\left(F(x_0) - F^* + \frac{6\tau_k}{\delta_k}T_0 + \frac{3\rho_Q\tilde{\sigma}^2}{2L_Q}\sum_{t=1}^k \frac{\tau_t}{\Theta_t}\right) \\ &\leq \Theta_k\left(F(x_0) - F^* + \frac{6\tau_k}{\delta_k}T_0\right) + \frac{3\rho_Q\tilde{\sigma}^2}{2L_Q}. \end{aligned}$$

Then, note that

$$\begin{aligned} T_0 &= \frac{5\delta_0}{12}(F(x_0) - F^*) + \frac{\mu}{2}\|x_0 - x^*\|^2 + \frac{5\delta_0}{24L_Q n}\sum_{i=1}^n \frac{1}{q_i n}\|u_0^i - u_*^i\|^2 \\ &\leq \frac{5\delta_0}{12}(F(x_0) - F^*) + \frac{\mu}{2}\|x_0 - x^*\|^2 + \frac{5\delta_0}{12}(F(x_0) - F^*), \end{aligned}$$

where the inequality comes from Theorem 2.1.5 of (Nesterov, 2004) and the definition of the u_0^i's. Then, we conclude by noting that $5\tau \leq 1$, and that $\alpha \leq 3\mu$ and we use Lemma C.3. □

D.6. Proof of Corollary 4

Proof. We start by following similar steps as in the proof of Corollary 3 to study the convergence of the first phase with constant step size. We note that with the choice of η_k, we have $\delta_k = \tau_k$ for all k. Then, we apply Theorem 2 and obtain

$$\begin{aligned}\mathbb{E}\left[F(\hat{x}_k) - F^* + 3\mu\|x_k - x^*\|^2\right] &\leq \Theta_k \left(F(x_0) - F^* + 6T_0 + 18\rho_Q\tilde{\sigma}^2\eta\sum_{t=1}^{k}\frac{\tau_t}{\Theta_t}\right)\\ &\leq \Theta_k\left(F(x_0) - F^* + 6T_0\right) + 18\rho_Q\tilde{\sigma}^2\eta.\end{aligned}$$

Then, we use the same upper-bound on T_0 as in the proof of Corollary 3, giving us $6T_0 \leq 5\delta_0(F(x_0) - F^*) + 3\mu\|x_0 - x^*\|^2 \leq 7(F(x_0) - F^*)$ since $\delta_0 = \mu\eta \leq 1/5$, which is sufficient to conclude that

$$\mathbb{E}\left[F(\hat{x}_k) - F^* + 3\mu\|x_k - x^*\|^2\right] \leq 8\Theta_k\left(F(x_0) - F^*\right) + 18\rho_Q\eta\tilde{\sigma}^2. \tag{23}$$

Then, we restart the procedure. Since the convergence rate (23) applies for the first stage with a constant step size, the number of iterations to ensure the condition $\mathbb{E}[F(\hat{x}_k) - F^*] \leq 24\eta\rho_Q\tilde{\sigma}^2$ is upper bounded by K with

$$K = O\left(\left(n + \frac{L_Q}{\mu}\right)\log\left(\frac{F(x_0) - F^*}{\varepsilon}\right)\right).$$

Then, we restart the optimization procedure, assuming from now on that $\mathbb{E}[F(x_0) - F^*] \leq 24\eta\rho_Q\tilde{\sigma}^2$, with decreasing step sizes $\eta_k = \min\left(\frac{2}{\mu(k+2)}, \eta\right)$, Then, since $\delta_k = \mu\eta_k \leq \frac{1}{5n}$, we have that $\tau_k = \delta_k$ for all k, and Theorem 2 gives us—note that here $\Gamma_k = \Theta_k$—

$$\mathbb{E}\left[F(\hat{x}_k) - F^*\right] \leq \Gamma_k\left(F(x_0) - F^* + 6T_0 + 18\rho_Q\tilde{\sigma}^2\sum_{t=1}^{k}\frac{\eta_t\delta_t}{\Gamma_t}\right) \quad \text{with} \quad \Gamma_k = \prod_{t=1}^{k}(1 - \delta_t).$$

Then, as noted in the proof of Corollary 4, we have $6T_0 \leq 7(F(x_0) - F^*)$. Then, after taking the expectation with respect to the output of the first stage,

$$\begin{aligned}\mathbb{E}\left[F(\hat{x}_k) - F^*\right] &\leq \Gamma_k\left(8\mathbb{E}[F(x_0) - F^*] + 18\rho_Q\tilde{\sigma}^2\sum_{t=1}^{k}\frac{\eta_t\delta_t}{\Gamma_t}\right)\\ &\leq \Gamma_k\left(192\rho_Q\eta\tilde{\sigma}^2 + 18\rho_Q\tilde{\sigma}^2\sum_{t=1}^{k}\frac{\eta_t\delta_t}{\Gamma_t}\right).\end{aligned}$$

Denote now by k_0 the largest index such that $\frac{2}{\mu(k_0+2)} \geq \eta$ and thus $k_0 = \lceil 2/(\mu\eta) - 2\rceil$. Then, according to Lemma C.4, for $k \geq k_0$,

$$\begin{aligned}\mathbb{E}\left[F(\hat{x}_k) - F^*\right] &\leq \Gamma_k\left(192\rho_Q\eta\tilde{\sigma}^2 + 18\rho_Q\eta\tilde{\sigma}^2\sum_{t=1}^{k_0-1}\frac{\delta_t}{\Gamma_t} + 18\rho_Q\tilde{\sigma}^2\sum_{t=k_0}^{k}\frac{2\delta_t}{\mu\Gamma_t(t+2)}\right)\\ &\leq \frac{k_0(k_0+1)}{(k+1)(k+2)}\left(\Gamma_{k_0-1}192\rho_Q\eta\tilde{\sigma}^2 + 18\eta\rho_Q\tilde{\sigma}^2\Gamma_{k_0-1}\sum_{t=1}^{k_0-1}\frac{\delta_t}{\Gamma_t}\right) + 36\rho_Q\tilde{\sigma}^2\sum_{t=k_0}^{k}\frac{\delta_t\Gamma_k}{\mu\Gamma_t(t+2)}\\ &\leq \frac{k_0(k_0+1)}{(k+1)(k+2)}192\eta\rho_Q\tilde{\sigma}^2 + 36\rho_Q\tilde{\sigma}^2\sum_{t=k_0}^{k}\frac{(t+1)(t+2)}{\mu(k+1)(k+2)(t+2)^2}\\ &\leq \frac{k_0\eta}{k+2}192\rho_Q\tilde{\sigma}^2 + \frac{36\rho_Q\tilde{\sigma}^2}{\mu(k+2)} = O\left(\frac{\rho_Q\tilde{\sigma}^2}{\mu k}\right),\end{aligned}$$

which gives the desired complexity.

$\square$

D.7. Proof of Theorem 3

Proof. First, the minimizer v_k of the quadratic surrogate d_k may be written as

$$
\begin{aligned}
v_k &= \frac{(1-\delta_k)\gamma_{k-1}}{\gamma_k} v_{k-1} + \frac{\mu\delta_k}{\gamma_k} y_{k-1} - \frac{\delta_k}{\gamma_k}\tilde{g}_k \\
&= y_{k-1} + \frac{(1-\delta_k)\gamma_{k-1}}{\gamma_k}(v_{k-1} - y_{k-1}) - \frac{\delta_k}{\gamma_k}\tilde{g}_k.
\end{aligned}
$$

Then, we characterize the quantity d_k^*:

$$
\begin{aligned}
d_k^* &= d_k(y_{k-1}) - \frac{\gamma_k}{2}\|v_k - y_{k-1}\|^2 \\
&= (1-\delta_k)d_{k-1}(y_{k-1}) + \delta_k l_k(y_{k-1}) - \frac{\gamma_k}{2}\|v_k - y_{k-1}\|^2 \\
&= (1-\delta_k)\left(d_{k-1}^* + \frac{\gamma_{k-1}}{2}\|y_{k-1} - v_{k-1}\|^2\right) + \delta_k l_k(y_{k-1}) - \frac{\gamma_k}{2}\|v_k - y_{k-1}\|^2 \\
&= (1-\delta_k)d_{k-1}^* + \left(\frac{\gamma_{k-1}(1-\delta_k)(\gamma_k - (1-\delta_k)\gamma_{k-1})}{2\gamma_k}\right)\|y_{k-1} - v_{k-1}\|^2 + \delta_k l_k(y_{k-1}) \\
&\qquad - \frac{\delta_k^2}{2\gamma_k}\|\tilde{g}_k\|^2 + \frac{\delta_k(1-\delta_k)\gamma_{k-1}}{\gamma_k}\tilde{g}_k^\top(v_{k-1} - y_{k-1}) \\
&\geq (1-\delta_k)d_{k-1}^* + \delta_k l_k(y_{k-1}) - \frac{\delta_k^2}{2\gamma_k}\|\tilde{g}_k\|^2 + \frac{\delta_k(1-\delta_k)\gamma_{k-1}}{\gamma_k}\tilde{g}_k^\top(v_{k-1} - y_{k-1}).
\end{aligned}
$$

Assuming by induction that $\mathbb{E}[d_{k-1}^*] \geq \mathbb{E}[F(x_{k-1})] - \xi_{k-1}$ for some $\xi_{k-1} \geq 0$, we have after taking expectation

$$
\mathbb{E}[d_k^*] \geq (1-\delta_k)(\mathbb{E}[F(x_{k-1})] - \xi_{k-1}) + \delta_k\mathbb{E}[l_k(y_{k-1})] - \frac{\delta_k^2}{2\gamma_k}\mathbb{E}\|\tilde{g}_k\|^2 + \frac{\delta_k(1-\delta_k)\gamma_{k-1}}{\gamma_k}\mathbb{E}[\tilde{g}_k^\top(v_{k-1} - y_{k-1})].
$$

Then, note that $\mathbb{E}[F(x_{k-1})] \geq \mathbb{E}[l_k(x_{k-1})] \geq \mathbb{E}[l_k(y_{k-1})] + \mathbb{E}[\tilde{g}_k^\top(x_{k-1} - y_{k-1})]$, and

$$
\mathbb{E}[d_k^*] \geq \mathbb{E}[l_k(y_{k-1})] - (1-\delta_k)\xi_{k-1} - \frac{\delta_k^2}{2\gamma_k}\mathbb{E}\|\tilde{g}_k\|^2 + (1-\delta_k)\mathbb{E}\left[\tilde{g}_k^\top\left(\frac{\delta_k\gamma_{k-1}}{\gamma_k}(v_{k-1} - y_{k-1}) + (x_{k-1} - y_{k-1})\right)\right].
$$

By Lemma 1, we can show that the last term is equal to zero, and we are left with

$$
\mathbb{E}[d_k^*] \geq \mathbb{E}[l_k(y_{k-1})] - (1-\delta_k)\xi_{k-1} - \frac{\delta_k^2}{2\gamma_k}\mathbb{E}\|\tilde{g}_k\|^2.
$$

We may then use Lemma 2, which gives us

$$
\begin{aligned}
\mathbb{E}[d_k^*] &\geq \mathbb{E}[F(x_k)] - (1-\delta_k)\xi_{k-1} - \eta_k\sigma_k^2 + \left(\eta_k - \frac{L\eta_k^2}{2} - \frac{\delta_k^2}{2\gamma_k}\right)\mathbb{E}\|\tilde{g}_k\|^2 \\
&\geq \mathbb{E}[F(x_k)] - \xi_k \quad \text{with} \quad \xi_k = (1-\delta_k)\xi_{k-1} + \eta_k\sigma_k^2,
\end{aligned}
$$

where we used the fact that $\eta_k \leq 1/L$ and $\delta_k = \sqrt{\gamma_k\eta_k}$.

It remains to choose $d_0^* = F(x_0)$ and $\xi_0 = 0$ to initialize the induction at $k = 0$ and we conclude that

$$
\mathbb{E}\left[F(x_k) - F^* + \frac{\gamma_k}{2}\|v_k - x^*\|^2\right] \leq \mathbb{E}[d_k(x^*) - F^*] + \xi_k \leq \Gamma_k(d_0(x^*) - F^*) + \xi_k,
$$

which gives us the desired result when noticing that $\xi_k = \Gamma_k\sum_{t=1}^{k}\frac{\eta_t\sigma_t^2}{\Gamma_t}$. $\square$

D.8. Proof of Lemma 1

Proof. Let us assume that the relation $y_{k-1} = \theta_{k-1}x_{k-1} + (1-\theta_{k-1})v_{k-1}$ holds and let us show that it also holds for y_k. Since the estimate sequences d_k are quadratic functions, we have

$$\begin{aligned}
v_k &= (1-\delta_k)\frac{\gamma_{k-1}}{\gamma_k}v_{k-1} + \frac{\mu\delta_k}{\gamma_k}y_{k-1} - \frac{\delta_k}{\gamma_k}(g_k + \psi'(x_k)) \\
&= (1-\delta_k)\frac{\gamma_{k-1}}{\gamma_k}v_{k-1} + \frac{\mu\delta_k}{\gamma_k}y_{k-1} - \frac{\delta_k}{\gamma_k\eta_k}(y_{k-1} - x_k) \\
&= (1-\delta_k)\frac{\gamma_{k-1}}{\gamma_k(1-\theta_{k-1})}(y_{k-1} - \theta_{k-1}x_{k-1}) + \frac{\mu\delta_k}{\gamma_k}y_{k-1} - \frac{\delta_k}{\gamma_k\eta_k}(y_{k-1} - x_k) \\
&= (1-\delta_k)\frac{\gamma_{k-1}}{\gamma_k(1-\theta_{k-1})}(y_{k-1} - \theta_{k-1}x_{k-1}) + \frac{\mu\delta_k}{\gamma_k}y_{k-1} - \frac{1}{\delta_k}(y_{k-1} - x_k) \\
&= \left(\frac{(1-\delta_k)\gamma_{k-1}}{\gamma_k(1-\theta_{k-1})} + \frac{\mu\delta_k}{\gamma_k} - \frac{1}{\delta_k}\right)y_{k-1} - \frac{(1-\delta_k)\gamma_{k-1}\theta_{k-1}}{\gamma_k(1-\theta_{k-1})}x_{k-1} + \frac{1}{\delta_k}x_k \\
&= \left(1 + \frac{(1-\delta_k)\gamma_{k-1}\theta_{k-1}}{\gamma_k(1-\theta_{k-1})} - \frac{1}{\delta_k}\right)y_{k-1} - \frac{(1-\delta_k)\gamma_{k-1}\theta_{k-1}}{\gamma_k(1-\theta_{k-1})}x_{k-1} + \frac{1}{\delta_k}x_k.
\end{aligned}$$

Then note that $1-\theta_{k-1} = \frac{\delta_k\gamma_{k-1}}{\gamma_{k-1}+\delta_k\mu}$ and thus, $\frac{\gamma_{k-1}\theta_{k-1}}{\gamma_k(1-\theta_{k-1})} = \frac{1}{\delta_k}$, and

$$v_k = x_{k-1} + \frac{1}{\delta_k}(x_k - x_{k-1}).$$

Then, we note that $x_k - x_{k-1} = \frac{\delta_k}{1-\delta_k}(v_k - x_k)$ and we are left with

$$y_k = x_k + \beta_k(x_k - x_{k-1}) = \frac{\beta_k\delta_k}{1-\delta_k}v_k + \left(1 - \frac{\beta_k\delta_k}{1-\delta_k}\right)x_k.$$

Then, it is easy to show that

$$\beta_k = \frac{(1-\delta_k)\delta_{k+1}\gamma_k}{\delta_k(\gamma_{k+1} + \delta_{k+1}\gamma_k)} = \frac{(1-\delta_k)\delta_{k+1}\gamma_k}{\delta_k(\gamma_k + \delta_{k+1}\mu)} = \frac{(1-\delta_k)(1-\theta_k)}{\delta_k},$$

which allows us to conclude that $y_k = \theta_k x_k + (1-\theta_k)v_k$ since the relation holds trivially for $k = 0$. □

D.9. Proof of Lemma 2

Proof.

$$\begin{aligned}
\mathbb{E}[F(x_k)] &= \mathbb{E}[f(x_k) + \psi(x_k)] \\
&\leq \mathbb{E}\left[f(y_{k-1}) + \nabla f(y_{k-1})^\top(x_k - y_{k-1}) + \frac{L}{2}\|x_k - y_{k-1}\|^2 + \psi(x_k)\right] \\
&= \mathbb{E}\left[f(y_{k-1}) + g_k^\top(x_k - y_{k-1}) + \frac{L}{2}\|x_k - y_{k-1}\|^2 + \psi(x_k)\right] + \mathbb{E}\left[(\nabla f(y_{k-1}) - g_k)^\top(x_k - y_{k-1})\right] \\
&= \mathbb{E}\left[f(y_{k-1}) + g_k^\top(x_k - y_{k-1}) + \frac{L}{2}\|x_k - y_{k-1}\|^2 + \psi(x_k)\right] + \mathbb{E}\left[(\nabla f(y_{k-1}) - g_k)^\top x_k\right] \\
&= \mathbb{E}\left[f(y_{k-1}) + g_k^\top(x_k - y_{k-1}) + \frac{L}{2}\|x_k - y_{k-1}\|^2 + \psi(x_k)\right] + \mathbb{E}\left[(\nabla f(y_{k-1}) - g_k)^\top(x_k - w_{k-1})\right] \\
&\leq \mathbb{E}\left[f(y_{k-1}) + g_k^\top(x_k - y_{k-1}) + \frac{L}{2}\|x_k - y_{k-1}\|^2 + \psi(x_k)\right] + \mathbb{E}\left[\|\nabla f(y_{k-1}) - g_k\|\|x_k - w_{k-1}\|\right] \\
&\leq \mathbb{E}\left[f(y_{k-1}) + g_k^\top(x_k - y_{k-1}) + \frac{L}{2}\|x_k - y_{k-1}\|^2 + \psi(x_k)\right] + \mathbb{E}\left[\eta_k\|\nabla f(y_{k-1}) - g_k\|^2\right] \\
&= \mathbb{E}\left[l_k(y_{k-1}) + \tilde{g}_k^\top(x_k - y_{k-1}) + \frac{L}{2}\|x_k - y_{k-1}\|^2\right] + \eta_k\sigma_k^2, \\
&\leq \mathbb{E}\left[l_k(y_{k-1})\right] + \left(\frac{L\eta_k^2}{2} - \eta_k\right)\mathbb{E}\left[\|\tilde{g}_k\|^2\right] + \eta_k\sigma_k^2,
\end{aligned}$$

where $w_{k-1} = \text{Prox}_{\eta_k \psi}[y_{k-1} - \eta_k \nabla f(y_{k-1})]$. The first inequality is due to the L-smoothness of f (Lemma C.1); then, the next three relations exploit the fact that $\mathbb{E}[(\nabla f(y_{k-1}) - g_k)^\top z = 0$ for all z that is deterministic (which is the case for y_{k-1} and w_{k-1}); the second inequality uses the non-expansiveness of the proximal operator. Then, we use the fact that $x_k = y_{k-1} - \eta_k \tilde{g}_k$. ☐

D.10. Proof of Corollary 6

Proof. The proof is similar to that of Corollary 2 for unaccelerated SGD. The first stage with constant step-size requires $O\left(\sqrt{\frac{L}{\mu}} \log\left(\frac{F(x_0)-F^*}{\varepsilon}\right)\right)$ iterations. Then, we restart the optimization procedure, and assume that $\mathbb{E}\left[F(x_0) - F^* + \frac{\mu}{2}\|x^* - x_0\|^2\right] \leq \frac{2\sigma^2}{\sqrt{\mu L}}$. With the choice of parameters, we have $\gamma_k = \mu$ and $\delta_k = \sqrt{\gamma_k \eta_k} = \min\left(\sqrt{\frac{\mu}{L}}, \frac{2}{k+2}\right)$. We may then apply Theorem 3 where the value of Γ_k is given by Lemma C.4. This yields for $k \geq k_0 = \left\lceil 2\sqrt{\frac{L}{\mu}} - 2 \right\rceil$,

$$
\begin{aligned}
\mathbb{E}[F(x_k) - F^*] &\leq \Gamma_k \left(\mathbb{E}\left[F(x_0) - F^* + \frac{\mu}{2}\|x_0 - x^*\|^2\right] + \sigma^2 \sum_{t=1}^{k} \frac{\eta_t}{\Gamma_t} \right) \\
&\leq \Gamma_k \left(\frac{2\sigma^2}{\sqrt{\mu L}} + \frac{\sigma^2}{L} \sum_{t=1}^{k_0-1} \frac{1}{\Gamma_t} + \sigma^2 \sum_{t=k_0}^{k} \frac{4}{\Gamma_t \mu (t+2)^2} \right) \\
&= \frac{k_0(k_0+1)}{(k+1)(k+2)} \left(\Gamma_{k_0-1} \frac{2\sigma^2}{\sqrt{\mu L}} + \frac{\sigma^2}{L} \Gamma_{k_0-1} \sum_{t=1}^{k_0-1} \frac{1}{\Gamma_t} \right) + \sigma^2 \sum_{t=k_0}^{k} \frac{4\Gamma_k}{\Gamma_t \mu (t+2)^2} \\
&= \frac{k_0(k_0+1)}{(k+1)(k+2)} \left(\Gamma_{k_0-1} \frac{2\sigma^2}{\sqrt{\mu L}} + (1 - \Gamma_{k_0-1}) \frac{\sigma^2}{\sqrt{\mu L}} \right) + \sigma^2 \sum_{t=k_0}^{k} \frac{4\Gamma_k}{\Gamma_t \mu (t+2)^2} \\
&\leq \frac{k_0(k_0+1)}{(k+1)(k+2)} \frac{2\sigma^2}{\sqrt{\mu L}} + \sigma^2 \frac{1}{(k+1)(k+2)} \left(\sum_{t=k_0+1}^{k} \frac{4(t+1)(t+2)}{\mu (t+2)^2} \right) \\
&\leq \frac{k_0}{(k+1)(k+2)} \frac{4\sigma^2}{\mu} + \frac{4\sigma^2}{\mu(k+2)} \leq \frac{8\sigma^2}{\mu(k+2)},
\end{aligned}
$$

where we use Lemmas C.4 and C.5. This leads to the desired iteration complexity. ☐

D.11. Proof of Proposition 4

Proof.

$$
\begin{aligned}
\sigma_k^2 &= \mathbb{E}\left\| \frac{1}{q_{i_k} n} \left(\tilde{\nabla} f_{i_k}(y_{k-1}) - \tilde{\nabla} f_{i_k}(\tilde{x}_{k-1}) \right) + \tilde{\nabla} f(\tilde{x}_{k-1}) - \nabla f(y_{k-1}) \right\|^2 \\
&= \mathbb{E}\left\| \frac{1}{q_{i_k} n} \left(\nabla f_{i_k}(y_{k-1}) + \zeta_k - \zeta_k' - \nabla f_{i_k}(\tilde{x}_{k-1}) \right) + \nabla f(\tilde{x}_{k-1}) + \bar{\zeta}_{k-1} - \nabla f(y_{k-1}) \right\|^2, \\
&\leq \mathbb{E}\left\| \frac{1}{q_{i_k} n} \left(\nabla f_{i_k}(y_{k-1}) - \nabla f_{i_k}(\tilde{x}_{k-1}) \right) + \nabla f(\tilde{x}_{k-1}) + \bar{\zeta}_{k-1} - \nabla f(y_{k-1}) \right\|^2 + 2\rho_Q \tilde{\sigma}^2,
\end{aligned}
$$

where ζ_k and ζ_k' are perturbations drawn at iteration k, and $\bar{\zeta}_{k-1}$ was drawn last time $\tilde{x}_{k-1}$ was updated. Then, by noticing that for any deterministic quantity Y and random variable X, we have $\mathbb{E}[\|X - \mathbb{E}[X] - Y\|^2] \leq \mathbb{E}[\|X\|^2] + \|Y\|^2$, taking expectation with respect to the index $i_k \sim Q$ and conditioning on $\mathcal{F}_{k-1}$, we have

$$\begin{aligned}
\sigma_k^2 &\leq \mathbb{E}\left\|\frac{1}{q_{i_k} n}\left(\nabla f_{i_k}(y_{k-1}) - \nabla f_{i_k}(\tilde{x}_{k-1})\right)\right\|^2 + \mathbb{E}[\|\bar{\zeta}_{k-1}\|^2] + 2\rho_Q\tilde{\sigma}^2 \\
&\leq \frac{1}{n}\sum_{i=1}^n \frac{1}{q_i n}\mathbb{E}\left\|\nabla f_i(y_{k-1}) - \nabla f_i(\tilde{x}_{k-1})\right\|^2 + 3\rho_Q\tilde{\sigma}^2 \\
&\leq \frac{1}{n}\sum_{i=1}^n \frac{2L_i}{q_i n}\mathbb{E}\left[f_i(\tilde{x}_{k-1}) - f_i(y_{k-1}) - \nabla f_i(y_{k-1})^\top(\tilde{x}_{k-1} - y_{k-1})\right] + 3\rho_Q\tilde{\sigma}^2 \\
&\leq \frac{1}{n}\sum_{i=1}^n 2L_Q\mathbb{E}\left[f_i(\tilde{x}_{k-1}) - f_i(y_{k-1}) - \nabla f_i(y_{k-1})^\top(\tilde{x}_{k-1} - y_{k-1})\right] + 3\rho_Q\tilde{\sigma}^2 \\
&= 2L_Q\mathbb{E}\left[f(\tilde{x}_{k-1}) - f(y_{k-1}) - \nabla f(y_{k-1})^\top(\tilde{x}_{k-1} - y_{k-1})\right] + 3\rho_Q\tilde{\sigma}^2 \\
&= 2L_Q\mathbb{E}\left[f(\tilde{x}_{k-1}) - f(y_{k-1}) - g_k^\top(\tilde{x}_{k-1} - y_{k-1})\right] + 3\rho_Q\tilde{\sigma}^2,
\end{aligned} \tag{24}$$

where the second inequality uses the upper-bound $\mathbb{E}[\|\bar{\zeta}\|^2] = \frac{\sigma^2}{n} \leq \rho_Q\sigma^2$, and the third one uses Theorem 2.1.5 in (Nesterov, 2004). □

D.12. Proof of Lemma 3

Proof. We can show that Lemma 2 still holds and thus,

$$\begin{aligned}
\mathbb{E}[F(x_k)] &\leq \mathbb{E}\left[l_k(y_{k-1})\right] + \left(\frac{L\eta_k^2}{2} - \eta_k\right)\mathbb{E}\left[\|\tilde{g}_k\|^2\right] + \eta_k\sigma_k^2. \\
&\leq \mathbb{E}\left[l_k(y_{k-1}) + a_k f(\tilde{x}_{k-1}) - a_k f(y_{k-1}) + a_k g_k^\top(y_{k-1} - \tilde{x}_{k-1})\right] \\
&\quad + \mathbb{E}\left[\left(\frac{L\eta_k^2}{2} - \eta_k\right)\|\tilde{g}_k\|^2\right] + 3\rho_Q\eta_k\tilde{\sigma}^2,
\end{aligned}$$

Note also that

$$\begin{aligned}
l_k(y_{k-1}) + f(\tilde{x}_{k-1}) - f(y_{k-1}) &= \psi(x_k) + \psi'(x_k)^\top(y_{k-1} - x_k) + f(\tilde{x}_{k-1}) \\
&\leq \psi(\tilde{x}_{k-1}) - \psi'(x_k)^\top(\tilde{x}_{k-1} - x_k) + \psi'(x_k)^\top(y_{k-1} - x_k) + f(\tilde{x}_{k-1}) \\
&= F(\tilde{x}_{k-1}) + \psi'(x_k)^\top(y_{k-1} - \tilde{x}_{k-1}).
\end{aligned}$$

Therefore, by noting that $l_k(y_{k-1}) + a_k f(\tilde{x}_{k-1}) - a_k f(y_{k-1}) \leq (1-a_k)l_k(y_{k-1}) + a_k F(\tilde{x}_{k-1}) + a_k\psi'(x_k)^\top(y_{k-1} - \tilde{x}_{k-1})$, we obtain the desired result. □

D.13. Proof of Theorem 4

Proof. Following similar steps as in the proof of Theorem 3, we have

$$d_k^* \geq (1-\delta_k)d_{k-1}^* + \delta_k l_k(y_{k-1}) - \frac{\delta_k^2}{2\gamma_k}\|\tilde{g}_k\|^2 + \frac{\delta_k(1-\delta_k)\gamma_{k-1}}{\gamma_k}\tilde{g}_k^\top(v_{k-1} - y_{k-1}).$$

Assume now by induction that $\mathbb{E}[d_{k-1}^*] \geq \mathbb{E}[F(\tilde{x}_{k-1})] - \xi_{k-1}$ for some $\xi_{k-1} \geq 0$ and note that $\delta_k \leq \frac{1-a_k}{n}$ since $a_k = 2L_Q\eta_k \leq \frac{2}{3}$ and $\delta_k = \sqrt{\frac{5\eta_k\gamma_k}{3n}} \leq \frac{1}{3n} \leq \frac{1-a_k}{n}$. Then,

$$\begin{aligned}
\mathbb{E}[d_k^*] &\geq (1-\delta_k)(\mathbb{E}[F(\tilde{x}_{k-1})] - \xi_{k-1}) + \delta_k\mathbb{E}[l_k(y_{k-1})] - \frac{\delta_k^2}{2\gamma_k}\mathbb{E}[\|\tilde{g}_k\|^2] + \mathbb{E}\left[\tilde{g}_k^\top\left(\frac{\delta_k(1-\delta_k)\gamma_{k-1}}{\gamma_k}(v_{k-1} - y_{k-1})\right)\right] \\
&\geq \left(1 - \frac{1-a_k}{n}\right)\mathbb{E}[F(\tilde{x}_{k-1})] + \left(\frac{1-a_k}{n} - \delta_k\right)\mathbb{E}[F(\tilde{x}_{k-1})] + \delta_k\mathbb{E}[l_k(y_{k-1})] - \frac{\delta_k^2}{2\gamma_k}\mathbb{E}[\|\tilde{g}_k\|^2] \\
&\quad + \mathbb{E}\left[\tilde{g}_k^\top\left(\frac{\delta_k(1-\delta_k)\gamma_{k-1}}{\gamma_k}(v_{k-1} - y_{k-1})\right)\right] - (1-\delta_k)\xi_{k-1}.
\end{aligned}$$

Note that

$$\mathbb{E}[F(\tilde{x}_{k-1})] \geq \mathbb{E}[l_k(\tilde{x}_{k-1})] \geq \mathbb{E}[l_k(y_{k-1})] + \mathbb{E}[\tilde{g}_k^\top(\tilde{x}_{k-1} - y_{k-1})].$$

Then,

$$\mathbb{E}[d_k^*] \geq \left(1 - \frac{1-a_k}{n}\right)\mathbb{E}[F(\tilde{x}_{k-1})] + \frac{1-a_k}{n}\mathbb{E}[l_k(y_{k-1})] - \frac{\delta_k^2}{2\gamma_k}\mathbb{E}[\|\tilde{g}_k\|^2]$$
$$+ \mathbb{E}\left[\tilde{g}_k^\top\left(\frac{\delta_k(1-\delta_k)\gamma_{k-1}}{\gamma_k}(v_{k-1} - y_{k-1}) + \left(\frac{1-a_k}{n} - \delta_k\right)(\tilde{x}_{k-1} - y_{k-1})\right)\right] - (1-\delta_k)\xi_{k-1}.$$

We may now use Lemma 3, which gives us

$$\mathbb{E}[d_k^*] \geq \left(1 - \frac{1}{n}\right)\mathbb{E}[F(\tilde{x}_{k-1})] + \frac{1}{n}\mathbb{E}[F(x_k)] + \left(\frac{1}{n}\left(\eta_k - \frac{L\eta_k^2}{2}\right) - \frac{\delta_k^2}{2\gamma_k}\right)\mathbb{E}[\|\tilde{g}_k\|^2]$$
$$+ \mathbb{E}\left[\tilde{g}_k^\top\left(\frac{\delta_k(1-\delta_k)\gamma_{k-1}}{\gamma_k}(v_{k-1} - y_{k-1}) + \left(\frac{1}{n} - \delta_k\right)(\tilde{x}_{k-1} - y_{k-1})\right)\right] - \xi_k, \quad (25)$$

with $\xi_k = (1-\delta_k)\xi_{k-1} + \frac{3\rho_Q\eta_k\tilde{\sigma}^2}{n}$. Then, since $\delta_k = \sqrt{\frac{5\eta_k\gamma_k}{3n}}$ and $\eta_k \leq \frac{1}{3L_Q} \leq \frac{1}{3L}$,

$$\frac{1}{n}\left(\eta_k - \frac{L\eta_k^2}{2}\right) - \frac{\delta_k^2}{2\gamma_k} \geq \frac{5\eta_k}{6n} - \frac{\delta_k^2}{2\gamma_k} = 0,$$

and the term in (25) involving $\|\tilde{g}_k\|^2$ may disappear. Similarly, we have

$$\frac{\delta_k(1-\delta_k)\gamma_{k-1}}{\delta_k(1-\delta_k)\gamma_{k-1} + \gamma_k/n - \delta_k\gamma_k} = \frac{\delta_k\gamma_k - \delta_k^2\mu}{\gamma_k/n - \delta_k^2\mu} = \frac{3n\delta_k^3/5\eta_k - \delta_k^2\mu}{3\delta_k^2/5\eta_k - \delta_k^2\mu} = \frac{3n - 5\mu\eta_k}{3 - 5\mu\eta_k} = \theta_k,$$

and the term in (25) that is linear in $\tilde{g}_k$ may disappear as well. Then, we are left with $\mathbb{E}[d_k^*] \geq \mathbb{E}[F(\tilde{x}_k)] - \xi_k$. Initializing the induction requires choosing $\xi_0 = 0$ and $d_0^* = F(x_0)$. Ultimately, we note that $\mathbb{E}[d_k(x^*) - F^*] \leq (1-\delta_k)\mathbb{E}[d_{k-1}(x^*) - F^*]$ for all $k \geq 1$, and

$$\mathbb{E}\left[F(\tilde{x}_k) - F^* + \frac{\gamma_k}{2}\|x^* - v_k\|^2\right] \leq \mathbb{E}[d_k(x^*) - F^*] + \xi_k \leq \Gamma_k\left(F(x_0) - F^* + \frac{\gamma_0}{2}\|x^* - x_0\|^2\right) + \xi_k,$$

and we obtain the desired result. □

D.14. Proof of Corollary 8

Proof. The proof is similar to that of Corollary 6 for accelerated SGD. The first stage with constant step-size η requires $O\left(\left(n + \sqrt{\frac{nL_Q}{\mu}}\right)\log\left(\frac{F(x_0)-F^*}{\varepsilon}\right)\right)$ iterations. Then, we restart the optimization procedure, and assume that $\mathbb{E}\left[F(x_0) - F^*\right] \leq B$ with $B = 3\rho_Q\tilde{\sigma}^2\sqrt{\eta/\mu n}$.

With the choice of parameters, we have $\gamma_k = \mu$ and $\delta_k = \sqrt{\frac{5\mu\eta_k}{3n}} = \min\left(\sqrt{\frac{5\mu\eta}{3n}}, \frac{2}{k+2}\right)$. We may then apply Theorem 4 where the value of Γ_k is given by Lemma C.4. This yields for $k \geq k_0 = \left\lceil \sqrt{\frac{12n}{5\mu\eta}} - 2 \right\rceil$,

$$\begin{aligned}
\mathbb{E}[F(x_k) - F^*] &\leq \Gamma_k \left(\mathbb{E}\left[F(x_0) - F^* + \frac{\mu}{2}\|x_0 - x^*\|^2 \right] + \frac{3\rho_Q \tilde{\sigma}^2}{n} \sum_{t=1}^{k} \frac{\eta_t}{\Gamma_t} \right) \\
&\leq \Gamma_k \left(2B + \frac{3\rho_Q \tilde{\sigma}^2 \eta}{n} \sum_{t=1}^{k_0-1} \frac{1}{\Gamma_t} + \frac{3\rho_Q \tilde{\sigma}^2}{n} \sum_{t=k_0}^{k} \frac{12n}{5\Gamma_t \mu (t+2)^2} \right) \\
&= \frac{k_0(k_0+1)}{(k+1)(k+2)} \left(\Gamma_{k_0-1} 2B + \frac{3\rho_Q \tilde{\sigma}^2 \eta}{n} \Gamma_{k_0-1} \sum_{t=1}^{k_0-1} \frac{1}{\Gamma_t} \right) + \frac{36\rho_Q \tilde{\sigma}^2}{5\mu} \sum_{t=k_0}^{k} \frac{\Gamma_k}{\Gamma_t (t+2)^2} \\
&= \frac{k_0(k_0+1)}{(k+1)(k+2)} \left(\Gamma_{k_0-1} 2B + (1 - \Gamma_{k_0-1}) \frac{3\rho_Q \tilde{\sigma}^2 \eta}{n\delta_{k_0}} \right) + \frac{36\rho_Q \tilde{\sigma}^2}{5\mu} \sum_{t=k_0}^{k} \frac{\Gamma_k}{\Gamma_t (t+2)^2} \\
&\leq \frac{2k_0(k_0+1)B}{(k+1)(k+2)} + \frac{8\rho_Q \tilde{\sigma}^2}{\mu(k+1)(k+2)} \left(\sum_{t=k_0+1}^{k} \frac{(t+1)(t+2)}{(t+2)^2} \right) \\
&\leq \frac{2k_0 B}{k+2} + \frac{8\rho_Q \tilde{\sigma}^2}{\mu(k+2)},
\end{aligned}$$

where we use Lemmas C.4 and C.5. Then, note that $k_0 B \leq 6\rho_Q \tilde{\sigma}^2/\mu$ and we obtain the right iteration complexity. $\square$