Estimate Sequences for Stochastic Composite Optimization: Variance Reduction, Acceleration, and Robustness to Noise

Andrei Kulunchakov, Julien Mairal

Estimate Sequences for Stochastic Composite Optimization: Variance Reduction, Acceleration, and Robustness to Noise

Andrei Kulunchakov
Inria*
andrei.kulunchakov@inria.fr

Julien Mairal
Inria*
julien.mairal.@inria.fr

January 25, 2019

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, we interpret a large class of stochastic optimization methods as procedures that iteratively minimize a surrogate of the objective. This point of view covers stochastic gradient descent (SGD), the variance-reduction approaches SAGA, SVRG, MISO, their proximal variants, and has several advantages: (i) we provide a simple generic proof of convergence for all of the aforementioned methods; (ii) we naturally obtain new algorithms with the same guarantees; (iii) 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 accelerated algorithms.

1 Introduction

We consider convex optimization problems of the form

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

where f is convex and L-smooth[1], and we call μ its strong convexity modulus with respect to the Euclidean norm.[2] The function ψ is convex lower semi-continuous and is not assumed to be necessarily differentiable. For instance, ψ may be the ℓ_1-norm, which is very popular in signal processing and machine learning for its sparsity-inducing properties [see Mairal et al., 2014, and references therein]; ψ may also be the extended-valued indicator function of a convex set $\mathcal{C}$ that takes the value $+\infty$ outside of $\mathcal{C}$ and 0 inside such that the previous setting encompasses constrained problems [see Hiriart-Urruty and Lemaréchal, 1996].

More specifically, we focus on stochastic objective functions, which are of utmost importance in machine learning, where f is an expectation or a finite sum of smooth functions

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

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. Whereas in machine learning formulations the explicit form of the data distribution is unknown, it is assumed that it is possible to draw from it random

*Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, 38000, France.

[1] A function is L-smooth when it is differentiable and its derivative is Lipschitz continuous with constant L.

[2] Then, $\mu = 0$ means that the function is convex but not strongly convex.

samples $\xi_1, \xi_2, \ldots$ Either an infinite number of such i.i.d. samples are available and the problem of interest is to minimize (1) with $f(x) = \mathbb{E}_\xi[\tilde{f}(x, \xi)]$, or one has access to a finite training set only, leading to the finite-sum setting on the right of (2), called empirical risk [Vapnik, 2000].

While the finite-sum setting is obviously a particular case of expectation with a discrete probability distribution, 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—which we assume is the case when f is an expectation—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]. Such a sublinear rate of convergence is notably achieved by stochastic gradient descent (SGD) algorithms or their variants [see Bottou et al., 2018, for a review].

Even though this pessimistic result applies to the general stochastic case, linear convergence rates can be obtained for the finite-sum setting [Schmidt et al., 2017]. Specifically, a large body of work in machine learning has led to many randomized incremental approaches such as SAG [Schmidt et al., 2017], SAGA [Defazio et al., 2014a], SVRG [Johnson and Zhang, 2013, Xiao and Zhang, 2014], SDCA [Shalev-Shwartz and Zhang, 2016], MISO [Mairal, 2015], Katyusha [Allen-Zhu, 2017], or the method of Lan and Zhou [2018]. These algorithms have about the same cost per-iteration as the stochastic gradient descent method, since they access only a single or two gradients $\nabla f_i(x)$ at each iteration, and they may achieve lower computational complexity than accelerated gradient descent methods [Nesterov, 1983, 2004, 2013, Beck and Teboulle, 2009] in expectation. A common interpretation is to see these algorithms as performing SGD steps with an estimate of the full gradient that has lower variance [Xiao and Zhang, 2014].

More precisely, accelerated gradient methods applied to the deterministic finite-sum problem when f is μ-strongly convex are able to provide an iterate x_k such that $F(x_k) - F^* \leq \varepsilon$ after $O(n\sqrt{L/\mu}\log(1/\varepsilon))$ evaluations of gradients $\nabla f_i(x)$ in the worst case; in contrast, variance-reduced stochastic optimization methods without acceleration achieve the same guarantee in expectation with complexity $O((n + L_Q/\mu)\log(1/\varepsilon))$, where $L_Q \geq L$ is the maximum Lipschitz constant of the gradients ∇f_i, or the average Lipschitz constant if a non-uniform sampling strategy Q is used. From a worst-case complexity point of view, the usefulness of these variance-reduced stochastic methods depend on n and on the discrepancy between L_Q and L. Indeed, when n is large enough, the complexity of these incremental approaches is simply $O(n\log(1/\varepsilon))$, which is independent of the condition number and always better than non-incremental first-order methods. Moreover, even though there is no guarantee that $L_Q \approx L$, large speed-ups over accelerated first-order methods have been reported on many classical machine learning datasets for incremental approaches [Defazio et al., 2014a, Schmidt et al., 2017], suggesting that L_Q is of the same order of magnitude as L in many practical cases. Note also that accelerated algorithms for finite sums have also been proposed by Shalev-Shwartz and Zhang [2016], Allen-Zhu [2017], Lan and Zhou [2018], Lin et al. [2018], which we will discuss later.

In this paper, we are interested in providing a unified view of stochastic optimization algorithms, with and without variance reduction, but we also want to investigate their *robustness* to random pertubations. Specifically, we may 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. For instance, perturbations may be injected during training in order to achieve better generalization on new test data [Srivastava et al., 2014], perform stable feature selection [Meinshausen and Bühlmann, 2010], improve the model robustness [Zheng et al., 2016], or for privacy-aware learning [Wainwright et al., 2012]. Each training point indexed by i is corrupted by a random perturbation ρ_i and the resulting function f may be written as

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

Whereas (3) is a finite sum of functions, we now assume that one has now only access to unbiased estimates of the gradients $\nabla f_i(x)$ due to the stochastic nature of f_i. Then, all the aforementioned variance-reduction methods do not apply anymore and the standard approach to address this problem is to ignore the finite-sum structure and use SGD or one of its variants. At each iteration, an estimate of the full gradient is obtained by randomly drawing an index $\hat{\imath}$ in $\{1, \ldots, n\}$ along with a perturbation. 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 the random sampling of the index $\hat{\imath}$ and $\tilde{\sigma}^2$ is due to the random data perturbation. In such a context, variance reduction consists of building gradient estimates with variance $\tilde{\sigma}^2$, which is potentially much smaller than σ^2. The SAGA and SVRG methods were adapted for such a purpose by Hofmann et al. [2015], though the resulting algorithms have non-zero asymptotic error; the MISO method was adapted by Bietti and Mairal [2017] at the cost of a memory overhead of $O(np)$, whereas other variants of SAGA and SVRG were proposed by Zheng and Kwok [2018] for linear models in machine learning.

The framework we adopt is that of estimate sequences introduced by Nesterov [2004], which consists of building iteratively a quadratic model of the objective. Typically, estimate sequences may be used to analyze the convergence of existing algorithms, but also to design new ones, in particular with acceleration. Our construction is however slightly different than the original one since it is based on stochastic estimates of the gradients, and some classical properties of estimate sequences are satisfied only approximately. We note that estimate sequences have been used before for stochastic optimization [Hu et al., 2009, Nitanda, 2014], but not in a generic fashion as we do and not for the same purpose. In summary, our construction leads to the following contributions:

- We revisit many stochastic optimization algorithms dealing with composite problems; in particular, we consider methods with variance reduction such as SVRG, SAGA, SDCA, or MISO. Interestingly, we show that all these algorithms admit variants that are adaptive to the strong convexity constant μ, when only a lower bound is available, a property that was not known for SDCA or MISO.

- We provide improvements to the previous algorithms by making them robust to stochastic perturbations. We analyze these approaches under a non-uniform sampling strategy $Q = \{q_1, \dots, q_n\}$ where q_i is the probability of drawing example i at each iteration. Typically, when the n gradients ∇f_i have different Lipschitz constants L_i, the uniform distribution Q yields complexities that depend on $L_Q = \max_i L_i$, whereas a non-uniform Q may yield $L_Q = \frac{1}{n}\sum_i L_i$. For strongly convex problems, the resulting worst-case iteration complexity for minimizing (3)—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 $\rho_Q = 1/(n \min q_i) \geq 1$ depends on the sampling strategy Q and $\rho_Q = 1$ for uniform distributions. The term on the left corresponds to the complexity of the variance-reduction methods for a deterministic objective 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 stochastic gradient descent for composite optimization applied to (3) has worst-case complexity $O(\sigma^2/\mu\varepsilon)$, with potentially $\sigma^2 \gg \tilde{\sigma}^2$. Note that the non-uniform sampling strategy potentially reduces L_Q and improves the left part, whereas it increases ρ_Q and degrades the dependency on the noise $\tilde{\sigma}^2$. Whereas non-uniform sampling strategies for incremental methods are now classical [Xiao and Zhang, 2014, Schmidt et al., 2015], the robustness to stochastic perturbations has not been studied for all these methods and existing approaches such as [Hofmann et al., 2015, Bietti and Mairal, 2017, Zheng and Kwok, 2018] have various limitations as discussed earlier.

- We show that our construction of estimate sequence naturally leads to an accelerated stochastic gradient method for composite optimization, similar in spirit to [Ghadimi and Lan, 2012, 2013, Hu et al., 2009], but slightly simpler. The resulting complexity in terms of gradient evaluations for μ-strongly convex objective is

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

  which, to the best of our knowledge, has only been achieved by Ghadimi and Lan [2013]. When the objective is convex, but non-strongly convex, we also provide a sublinear convergence rate for finite

horizon. Given a budget of K iterations, the algorithm returns an iterate x_K such that

$$\mathbb{E}[F(x_K) - F^*] \leq \frac{2L\|x_0 - x^*\|^2}{(K+1)^2} + \sigma\sqrt{\frac{8\|x_0 - x^*\|^2}{K+1}}, \tag{4}$$

which is also optimal for stochastic first-order optimization [Ghadimi and Lan, 2012].

- We design a new accelerated algorithm for finite sums based on the SVRG gradient estimator, with complexity, for μ-strongly convex functions,

$$O\left(\left(n + \sqrt{n\frac{L_Q}{\mu}}\right)\log\left(\frac{F(x_0) - F^*}{\varepsilon}\right)\right) + O\left(\frac{\rho_Q\tilde{\sigma}^2}{\mu\varepsilon}\right), \tag{5}$$

where the term on the left is the classical optimal complexity for finite sums [Arjevani and Shamir, 2016]. Note that when $\tilde{\sigma}^2 = 0$ (deterministic setting) we recover a similar complexity as Katyusha [Allen-Zhu, 2017]. When the problem is convex but not strongly convex, given a budget of K iterations that is large enough, the algorithm returns a solution x_K such that

$$\mathbb{E}[F(x_K) - F^*] \leq \frac{9nL_Q\|x_0 - x^*\|^2}{(K+1)^2} + \tilde{\sigma}\sqrt{\frac{12\rho_Q\|x_0 - x^*\|^2}{K+1}}, \tag{6}$$

where the term on the right is potentially better than (4) for large K when $\tilde{\sigma} \ll \sigma$ (see discussion above on full variance vs. variance due to small stochastic perturbations). When the objective is deterministic ($\tilde{\sigma} = 0$), the term (6) yields a complexity in $O(\sqrt{nL_Q}/\sqrt{\varepsilon})$, which is potentially better than the complexity $O(n\sqrt{L}/\sqrt{\varepsilon})$ of accelerated gradient descent, unless L is significantly smaller than L_Q.

This paper is organized as follows. Section 2 introduces the proposed framework based on estimate sequences; Section 3 is devoted to the convergence analysis; Section 4 presents a variant of SVRG with acceleration; Section 5 presents various experiments to compare the effectiveness of the proposed approaches, and Section 6 concludes the paper.

2 Proposed Framework Based on Stochastic Estimate Sequences

In this section, we present two generic stochastic optimization algorithms to address the composite problem (1). Then, we show their relation to variance-reduction methods.

2.1 A Classical Iteration Revisited

Consider an algorithm that performs the following updates:

$$x_k \leftarrow \mathrm{Prox}_{\eta_k\psi}\left[x_{k-1} - \eta_k g_k\right] \quad \text{with} \quad \mathbb{E}[g_k|\mathcal{F}_{k-1}] = \nabla f(x_{k-1}), \tag{A}$$

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

$$\mathrm{Prox}_{\eta\psi}[u] := \operatorname*{argmin}_{x\in\mathbb{R}^p}\left\{\eta\psi(x) + \frac{1}{2}\|x - u\|^2\right\}. \tag{7}$$

The iteration (A) is generic and encompasses many existing algorithms, which we review later. Key to our analysis, we are interested in a simple interpretation corresponding to the iterative minimization of strongly convex surrogate functions.

Interpretation with stochastic estimate sequence. Consider now the function

$$d_0(x) = d_0^* + \frac{\gamma_0}{2}\|x - x_0\|^2, \tag{8}$$

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 the following quadratic function d_k defined for $k \geq 1$ as

$$\begin{aligned} d_k(x) = (1-\delta_k)d_{k-1}(x) \\ + \delta_k \left(f(x_{k-1}) + g_k^\top (x - x_{k-1}) + \frac{\mu}{2}\|x - x_{k-1}\|^2 + \psi(x_k) + \psi'(x_k)^\top (x - x_k) \right), \end{aligned} \tag{9}$$

where δ_k, γ_k satisfy the system of equations

$$\delta_k = \eta_k \gamma_k \quad \text{and} \quad \gamma_k = (1-\delta_k)\gamma_{k-1} + \mu\delta_k, \tag{10}$$

and

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

We note that $\psi'(x_k)$ is a subgradient in $\partial\psi(x_k)$. By simply using the definition of the proximal operator (7) 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 generic form

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

The construction (9) 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 property of estimate sequences that we will nevertheless use is their ability do behave asymptotically as a lower bound of the objective function near the optimum. Indeed, we have

$$\mathbb{E}[d_k(x^*)] \leq (1-\delta_k)\mathbb{E}[d_{k-1}(x^*)] + \delta_k F^* \leq \Gamma_k d_0(x^*) + (1-\Gamma_k)F^*, \tag{11}$$

where $\Gamma_k = \prod_{t=1}^k (1-\delta_t)$ and $F^* = F(x^*)$. The first inequality comes from a strong convexity inequality since $\mathbb{E}[g_k^\top (x^* - x_{k-1})|\mathcal{F}_{k-1}] = \nabla f(x_{k-1})^\top (x^* - x_{k-1})$, and the second inequality is obtained by unrolling the relation obtained between $\mathbb{E}[d_k(x^*)]$ and $\mathbb{E}[d_{k-1}(x^*)]$. 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 descent), which uses the exact gradient $g_k = \nabla f(x_{k-1})$ leading to deterministic iterates $(x_k)_{k\geq 0}$ [Beck and Teboulle, 2009, Nesterov, 2013] or proximal variants of the stochastic gradient descent method to deal with a composite objective [see Lan, 2012, for instance]. Of interest for us, the variance-reduced stochastic optimization approaches SVRG [Xiao and Zhang, 2014] and SAGA [Defazio et al., 2014a] also follow the iteration (A) but with an unbiased gradient estimator whose variance reduces over time. Specifically, the basic form of these estimators is

$$g_k = \nabla f_{i_k}(x_{k-1}) - z_{k-1}^{i_k} + \bar{z}_{k-1} \quad \text{with} \quad \bar{z}_{k-1} = \frac{1}{n}\sum_{i=1}^n z_{k-1}^i, \tag{12}$$

where i_k is an index chosen uniformly in $\{1, \ldots, n\}$ at random, and each auxiliary variable z_k^i is equal to the 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 and SAGA are two different

approaches to build such positively correlated variables. 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 $\tilde{z}_k = \nabla f(\tilde{x}_k)$, which is thus $O(p)$. On the other hand, SAGA updates only $z_k^{i_k} = \nabla f_{i_k}(x_{k-1})$ at iteration k, such that $z_k^i = z_{k-1}^i$ 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., 2014a]. Note that variants with non-uniform sampling of the indices i_k have been proposed by Xiao and Zhang [2014], Schmidt et al. [2015].

2.2 Another Algorithm with a Different Estimate Sequence

In the previous section, we have interpreted the classical iteration (A) as the iterative minimization of the stochastic surrogate (9). Here, we show that a slightly different construction leads to a new algorithm. To obtain a lower bound, we have indeed used basic properties of the proximal operator to obtain a subgradient $\psi'(x_k)$ and we have exploited the following convexity inequality

$$\psi(x) \geq \psi(x_k) + \psi'(x_k)^\top (x - x_k).$$

Another natural choice to build a lower bound consists then of using directly $\psi(x)$ instead of $\psi(x_k) + \psi'(x_k)^\top (x - x_k)$, leading to the construction

$$d_k(x) = (1 - \delta_k) d_{k-1}(x) + \delta_k \left(f(x_{k-1}) + g_k^\top (x - x_{k-1}) + \frac{\mu}{2} \|x - x_{k-1}\|^2 + \psi(x) \right), \tag{13}$$

where x_{k-1} is assumed to be the minimizer of the composite function d_{k-1}, δ_k is defined as in Section 2.1, and x_k is a minimizer of d_k. To initialize the recursion, we define then d_0 as

$$d_0(x) = c_0 + \frac{\gamma_0}{2} \|x - \bar{x}_0\|^2 + \psi(x) \geq d_0^* + \frac{\gamma_0}{2} \|x - x_0\|^2,$$

with $x_0 = \mathrm{Prox}_{\psi/\gamma_0}[\bar{x}_0]$ is the minimizer of d_0 and $d_0^* = d_0(x_0) = c_0 + \frac{\gamma_0}{2} \|x_0 - \bar{x}_0\|^2 + \psi(x_0)$ is the minimum value of d_0; c_0 is left unspecified since it does not affect the algorithm. Typically, one may choose $\bar{x}_0$ to be a minimizer of ψ such that $x_0 = \bar{x}_0$. Unlike in the previous section, the surrogates d_k are not quadratic, but they remain γ_k-strongly convex. It is also easy to check that the relation (11) still holds.

The corresponding algorithm. It is also relatively easy to show that the iterative minimization of the stochastic lower bounds (13) leads to the following iterations

$$\bar{x}_k \leftarrow (1 - \mu\eta_k)\bar{x}_{k-1} + \mu\eta_k x_{k-1} - \eta_k g_k \quad \text{and} \quad x_k = \mathrm{Prox}_{\frac{\psi}{\gamma_k}}[\bar{x}_k] \quad \text{with} \quad \mathbb{E}[g_k | \mathcal{F}_{k-1}] = \nabla f(x_{k-1}). \tag{B}$$

As we will see, the convergence analysis for algorithm (A) also holds for algorithm (B) such that both variants enjoy similar theoretical properties. In one case, the function $\psi(x)$ appears explicitly, whereas a lower bound $\psi(x_k) + \psi'(x_k)^\top (x - x_k)$ is used in the other case. The introduction of the variable $\bar{x}_k$ allows us to write the surrogates d_k in the canonical form

$$d_k(x) = c_k + \frac{\gamma_k}{2} \|x - \bar{x}_k\|^2 + \psi(x) \geq d_k^* + \frac{\gamma_k}{2} \|x - x_k\|^2,$$

where c_k is constant and the inequality on the right is due to the strong convexity of d_k.

Relation to existing approaches. The approach (B) is related to several optimization methods. When the objective is a deterministic finite sum, the MISO algorithm [Mairal, 2015], one variant of SDCA [Shalev-Shwartz and Zhang, 2016], and Finito [Defazio et al., 2014b] adopt similar strategies and perform the update (B), even though they were derived from a significantly different point of view. For instance, SDCA is a

dual coordinate ascent approach, whereas MISO is explicitly derived from the iterative surrogate minimization we adopt in this paper. While the links between (B) and the previous approaches are not necessarily obvious when looking at the original description of these methods, it may be shown that they indeed perform such an update with a gradient estimator of the form (12) where $z_{i_k}^k = \nabla f_{i_k}(x_{k-1}) - \mu x_{k-1}$, where $\mu > 0$ is the strong convexity constant of the objective and $z_i^k = z_i^{k-1}$ if $i \neq i_k$. Whereas such estimator requires storing n gradients in general, the cost may be also reduced to $O(n)$ when dealing with a linear model in machine learning with a quadratic regularization function $\frac{\mu}{2}\|x\|^2$. Variants with non-uniform sampling for the index i_k appear also in the literature [Csiba et al., 2015, Bietti and Mairal, 2017].

2.3 Gradient Estimators and New Algorithms

In this paper, we consider the iterations (A) and (B) with the following gradient estimators that are variants of the ones above. For all of them, we define the variance σ_k to be

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

- **exact gradient**, with $g_k = \nabla f(x_{k-1})$, when the problem is deterministic ($\sigma_k = 0$);
- **stochastic gradient**, when we assume that g_k has bounded variance. Typically, when $f(x) = \mathbb{E}_\xi[\tilde{f}(x,\xi)]$, a data point ξ_k is drawn at iteration k and $g_k = \nabla \tilde{f}(x, \xi_k)$.
- **random-SVRG**: for finite sums, we consider a variant of the SVRG gradient estimator with non-uniform sampling and a random update of the anchor point $\tilde{x}_{k-1}$. Specifically, g_k is also an unbiased estimator of $\nabla f(x_{k-1})$, defined as

$$g_k = \frac{1}{q_{i_k} n}\left(\tilde{\nabla} f_{i_k}(x_{k-1}) - z_{k-1}^{i_k}\right) + \bar{z}_{k-1}, \tag{14}$$

  where i_k is sampled from a distribution $Q = \{q_1, \dots, q_n\}$ and $\tilde{\nabla}$ denotes that the gradient is perturbed by a zero-mean noise variable with variance $\tilde{\sigma}^2$. More precisely, if $f_i(x) = \mathbb{E}_\rho[\tilde{f}_i(x,\rho)]$ for all i, where ρ is a stochastic perturbation, instead of accessing $\nabla f_{i_k}(x_{k-1})$, we draw a perturbation ρ_k and observe

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

  where the perturbation ζ_k has zero mean given $\mathcal{F}_{k-1}$ and its variance is bounded by $\tilde{\sigma}^2$. When there is no perturbation, we simply have $\tilde{\nabla} = \nabla$ and $\zeta_k = 0$.

  Similar to the previous case, the variables z_k^i and $\bar{z}_k$ also correspond to possibly 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 [Xiao and Zhang, 2014] updates $\tilde{x}_k$ on a fixed schedule, we prefer to 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/SDCA or MISO. The use of this estimator with iteration (A) is illustrated in Algorithm 1. It is then easy to modify it to use variant (B) instead.

  In terms of memory, the random-SVRG gradient estimator requires to store an anchor point $\tilde{x}_{k-1}$ and the average gradients $\bar{z}_{k-1}$. 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 $z_{k-1}^{i_k} = \tilde{\nabla} f_{i_k}(\tilde{x}_{k-1})$ at iteration k, with the same perturbation for index i_k that was used to compute $\bar{z}_{k-1} = \frac{1}{n}\sum_{i=1}^n z_{k-1}^i$ when the anchor point was last updated. The overall cost is thus $O(n+p)$.

Algorithm 1 Variant (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 the distribution $Q = \{q_1, \ldots, q_n\}$;
  Compute the gradient estimator, possibly corrupted by random perturbations:

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

  Obtain the new iterate

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

  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 $\tilde{x}_k = \tilde{x}_{k-1}$ and $\bar{z}_k = \bar{z}_{k-1}$;
  Optional: Use the online averaging strategy using δ_k obtained from (10):

$$\hat{x}_k = (1-\tau_k)\hat{x}_{k-1} + \tau_k x_k \quad \text{with} \quad \tau_k = \min\left(\delta_k, \frac{1}{5n}\right);$$

end for
Output: x_K or $\hat{x}_K$ if averaging.

- **SAGA**: The estimator has a form similar to (14) 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 for $i = 1, \ldots, n$, while maintaining the relation $\bar{z}_k = \frac{1}{n}\sum_{i=1}^n z_k^i$. We consider variants such that each time a gradient $\nabla f_i(x)$ is computed, it is corrupted by a zero-mean random perturbation with variance $\tilde{\sigma}^2$. The procedure is described in Algorithm 2 for variant (A), with a more general estimator that encompasses SAGA/SDCA/MISO, as detailed next.

  Note that to deal with non-uniform sampling, we draw uniformly in $\{1, \ldots, n\}$ an index j_k for updating a variable $z_k^{j_k}$. When i_k is already drawn from a uniform distribution, we may choose instead $j_k = i_k$, which saves computations and does not affect the convergence results. The reason for using an additional index j_k in the non-uniform sampling case removes a difficulty in the convergence proof, a strategy also adopted by Schmidt et al. [2015] for a variant of SAGA with non-uniform sampling.

- **SDCA/MISO**: To put SAGA, MISO and SDCA under the same umbrella, we introduce a scalar β in $[0, \mu]$, which we will explain in the sequel, and a correcting term involving β that appears only when the sampling distribution Q is not uniform:

$$g_k = \frac{1}{q_{i_k} n}\left(\tilde{\nabla} f_{i_k}(x_{k-1}) - \beta x_{k-1} - z_{k-1}^{i_k}\right) + \bar{z}_{k-1} + \beta x_{k-1}. \tag{15}$$

  The resulting algorithm corresponds to a variant of SAGA when $\beta = 0$; when instead the gradient estimator is used in the context of variant (B), the choice $\beta = \mu$ then leads to MISO/SDCA-like procedures. The motivation for introducing the parameter β comes from regularized empirical risk minimization problems, where the functions f_i may have the form $f_i(x) = \phi(a_i^\top x) + \frac{\beta}{2}\|x\|^2$, where a_i in $\mathbb{R}^p$ is a data point; then, β is a lower bound on the strong convexity modulus, and $\nabla f_i(x) - \beta x$ is proportional to a_i, which is assumed to be already in memory. When there is no noise (meaning $\tilde{\sigma}^2 = 0$), storing the variables z_i^k then requires only n additional scalars.

Algorithm 2 Variant (A) with SAGA/SDCA/MISO estimator

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, $\gamma_0 \geq \mu$.
Initialization: $z_0^i = \tilde{\nabla} f_i(x_0) - \beta x_0$ for all $i = 1, \dots, n$ and $\bar{z}_0 = \frac{1}{n}\sum_{i=1}^n z_0^i$.
for $k = 1, \dots, K$ **do**
    Sample i_k according to the distribution $Q = \{q_1, \dots, q_n\}$;
    Compute the gradient estimator, possibly corrupted by random perturbations:

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

    Obtain the new iterate

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

    Draw j_k from the uniform distribution in $\{1, \dots, n\}$;
    Update the auxiliary variables

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

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

New Features. After having introduced our algorithms and before presenting the convergence analysis, we summarize here their new features.

- **robustness to noise**: As mentioned already in the introduction, we introduce mechanisms to deal with stochastic perturbations.
- **adaptivity to the strong convexity when $\tilde{\sigma} = 0$**: Algorithms 1 and 2 without averaging do not require knowing the strong convexity constant μ (MISO will simply need a lower-bound β, which is often trivial to obtain). As shown in the next section, no averaging simply leads to a slightly worse expected convergence rate.
- **new variants:** Whereas SVRG/SAGA were originally developed with the iterations (A) and SDCA or MISO in the context of (B), we show that these gradient estimators are compatible with (A) and (B).

3 Convergence Analysis and Robustness

We now present the convergence analysis of the algorithms described previously. In Section 3.1, we present a general convergence result. Then, we present specific results for the variance-reduction approaches in Section 3.2, including strategies to make them robust to stochastic noise. Acceleration will be discussed in the next section.

3.1 Generic Convergence Result Without Variance Reduction

Key to our complexity results, the following proposition gives a first relation between the quantity $F(x_k)$, the surrogate d_k, d_{k-1} and the variance σ_k of the gradient estimates.

Proposition 1 (Key relation). *For either variant (A) or (B), when using the construction of d_k from Sections 2.1 or 2.2, respectively, and assuming $\eta_k \leq 1/L$, we have for all $k \geq 1$,*

$$\delta_k(\mathbb{E}[F(x_k)] - F^*) + \mathbb{E}[d_k(x^*) - d_k^*] \leq (1 - \delta_k)\mathbb{E}[d_{k-1}(x^*) - d_{k-1}^*] + \eta_k \delta_k \sigma_k^2, \tag{16}$$

where F^* is the minimum of F, x^* is one of its minimizers, and $\sigma_k^2 = \mathbb{E}[\|g_k - \nabla f(x_{k-1})\|^2]$.

Proof. We first consider the variant (A) and later show how to modify the convergence proofs to accommodate the variant (B).

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

where the first inequality comes from Lemma 7—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 5. 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

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

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 (11) when $z = x^*$, and we obtain (16). It is also easy to see that the proof also works with variant (B). The convergence analysis is identical, except that we take w_{k-1} to be

$$w_{k-1} = \mathrm{Prox}_{\frac{\psi}{\gamma_k}}\left[(1 - \mu\eta_k)\bar{x}_{k-1} + \mu\eta_k x_{k-1} - \eta_k \nabla f(x_{k-1}) \right],$$

and the same result follows. ∎

Then, without making further assumption on σ_k, we have the following general convergence result, which is a direct consequence of the averaging Lemma 13, inspired by Ghadimi and Lan [2012], and presented in the appendix:

Theorem 1 (General convergence result). *Under the same assumptions as in Proposition 1, we have for all $k \geq 1$,*

$$\mathbb{E}[\delta_k (F(x_k) - F^*) + d_k(x^*) - d_k^*] \leq \Gamma_k \left(d_0(x^*) - d_0^* + \sum_{t=1}^k \frac{\delta_t \eta_t \sigma_t^2}{\Gamma_t} \right), \tag{17}$$

where $\Gamma_k = \prod_{t=1}^k (1 - \delta_t)$. Then, by using the averaging strategy of Lemma 13, which produces an iterate $\hat{x}_k$, we have

$$\mathbb{E}\left[F(\hat{x}_k) - F^* + d_k(x^*) - d_k^* \right] \leq \Gamma_k \left(F(x_0) - F^* + d_0(x^*) - d_0^* + \sum_{t=1}^k \frac{\delta_t \eta_t \sigma_t^2}{\Gamma_t} \right). \tag{18}$$

Theorem 1 allows us to recover convergence rates for various algorithms. Note that the effect of the averaging strategy is to remove the factor δ_k in front of $F(x_k) - F^*$ on the left part of (17), thus improving the convergence rate by a factor $1/\delta_k$. The price to pay is an additional constant term $F(x_0) - F^*$. For variant (A), the quantity $d_0(x^*) - d_0^*$ is equal to $\frac{\gamma_0}{2}\|x^* - x_0\|^2$, whereas it may be larger for (B). Indeed, we may simply say that $d_0(x^*) - d_0^* = \frac{\gamma_0}{2}\|x^* - x_0\|^2 + \psi(x^*) - \psi(x_0) - \psi'(x_0)^\top(x_0 - x^*)$ for variant (B), where $\psi'(x_0) = \gamma_0(x_0 - \bar{x}_0)$ is a subgradient in $\partial\psi(x_0)$.

As an illustration, we now provide various corollaries of the convergence result for variant (A) only. Note that the following lemma may further refine the upper-bound:

Lemma 1 (Auxiliary lemma for stochastic gradient descent iteration)**.**
Assume that there exists a point $\bar{x}_0$ such that $x_0 = Prox_{\eta_0\psi}[\bar{x}_0 - \eta_0 g_0]$ with $\mathbb{E}[g_0] = \nabla f(\bar{x}_0)$ and $\eta_0 \leq 1/L$. Then,

$$\mathbb{E}\left[F(x_0) + \frac{1}{2\eta_0}\|x_0 - x^*\|^2\right] \leq F^* + \frac{1}{2\eta_0}\|\bar{x}_0 - x^*\|^2 + \eta_0\sigma_0^2,$$

where σ_0^2 is the variance of g_0. Then, consider the iterates produced by Algorithm (A) with $\gamma_0 = \frac{1}{\eta_0}$. As a consequence, the iterates produced by Algorithm (A) satisfy

$$\mathbb{E}\left[F(\hat{x}_k) - F^* + \frac{\mu}{2}\|x_k - x^*\|^2\right] \leq \Gamma_k\left(\frac{1}{2\eta_0}\|\bar{x}_0 - x^*\|^2 + \eta_0\sigma_0^2 + \sum_{t=1}^{k}\frac{\delta_t\eta_t\sigma_t^2}{\Gamma_t}\right). \tag{19}$$

The purpose of the lemma is to replace the dependency in $F(x_0) - F^*$ by $\frac{L}{2}\|x_0 - x^*\|^2$ in the convergence rate (when $\eta_0 = 1/L$), at the price of one extra iteration. Whereas the latter naturally upper-bounds the former in the smooth case, we do not have $F(x_0) - F^* \leq \frac{L}{2}\|x_0 - x^*\|^2$ in the composite case, which makes result (19) slightly stronger than (18).

In the corollary below, we consider the stochastic setting showing that 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 11, which provides the convergence rate of $(\Gamma_k)_{k\geq 0}$ and uses also the relation $\Gamma_k \sum_{t=1}^{k} \frac{\delta_t}{\Gamma_t} = 1 - \Gamma_k \leq 1$ from Lemma 10 in the appendix, and Theorem 1.

Corollary 1 (**Proximal variants of SGD with constant step-size, $\mu > 0$**)**.**
Assume that f is μ-strongly convex, that the gradient estimates have constant variance $\sigma_k = \sigma$, and choose $\gamma_0 = \mu$ and $\eta_k = 1/L$ with Algorithm (A). Then,

$$\mathbb{E}\left[F(\hat{x}_k) - F^* + \frac{\mu}{2}\|x_k - x^*\|^2\right] \leq \left(1 - \frac{\mu}{L}\right)^k\left(F(x_0) - F^* + \frac{\mu}{2}\|x_0 - x^*\|^2\right) + \frac{\sigma^2}{L}. \tag{20}$$

We now show that it is possible to obtain converging algorithms by using decreasing step sizes. The proof of the following corollary is given in the appendix.

Corollary 2 (**Proximal variants of SGD with decreasing step-sizes with $\mu > 0$**)**.**
Assume that f is μ-strongly convex and that we target an accuracy ε smaller than $2\sigma^2/L$. First, use a constant step-size $\eta_k = 1/L$ with $\gamma_0 = \mu$ within Algorithm (A), leading to the convergence rate (20), until $\mathbb{E}[F(\hat{x}_k) - F^*] \leq 2\sigma^2/L$. Then, we 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 exactly when to start decreasing the step sizes in stochastic gradient descent algorithms is an open problem. The corollary simply supports the classical heuristic consisting of adopting a constant step size strategy long enough until the iterates oscillate without much progress, before decreasing the step sizes [see Bottou et al., 2018]. Next, we show that convergence to a noise dominated region of radius σ^2/L also holds when $\mu = 0$.

Corollary 3 (**Proximal variants of SGD with $\mu = 0$ and constant step size**).
Assume that f is convex, that the gradient estimates have constant variance $\sigma_k = \sigma$, and that for all $\eta_k = \eta \leq \frac{1}{L}$, choose $\gamma_0 = \frac{1}{\eta}$ with Algorithm (A). Then,

$$\mathbb{E}\left[F(\hat{x}_k) - F^*\right] \leq \frac{1}{2\eta k}\|x_0 - x^*\|^2 + \eta\sigma^2. \tag{21}$$

The proof of the previous corollary (i) considers the first iteration as part of Lemma 1 (the relation between x_1 and x_0 is the same as that of x_0 and $\bar{x}_0$ in the Lemma); (ii) uses Lemma 11 to obtain the rate of convergence of $\Gamma_k \leq \frac{1}{k+1}$ in (19), (ii) uses Lemma 10, which provides the simple relation $\Gamma_k \sum_{t=1}^{k} \frac{\delta_t}{\Gamma_t} = 1 - \Gamma_k$.

Corollary 4 (**Proximal variants of SGD with $\mu = 0$, finite horizon**). *Assume that f is convex, that the gradient estimates have constant variance $\sigma_k = \sigma$ and that we have a budget of K iterations for Algorithm (A). Choose a constant step size*

$$\eta_k = \min\left(\frac{1}{L}, \sqrt{\frac{T_0}{K\sigma^2}}\right) \quad \textit{with} \quad T_0 = \frac{1}{2}\|x_0 - x^*\|^2.$$

Then, with $\gamma_0 = 1/\eta$,

$$\mathbb{E}[F(\hat{x}_K) - F^*] \leq \frac{LT_0}{K} + 2\sigma\sqrt{\frac{T_0}{K}}. \tag{22}$$

This corollary is obtained by optimizing (21) with respect to η under the constraint $\eta \leq 1/L$. Considering both cases $\eta = 1/L$ and $\eta = \sqrt{T_0/K\sigma^2}$, it is easy to check that we have (22) in all cases. Whereas this last result is not a practical one since the step size depends on unknown quantities, it shows that our analysis is nevertheless able to recover the optimal noise-dependency in $O(\sigma\sqrt{T_0/K})$, [see Nemirovski et al., 2009]. Next, we focus on variance reduction mechanisms, which are able to improve the previous convergence rates by better exploiting the structure of the objective.

3.2 Faster Convergence with Variance Reduction

Stochastic variance-reduced gradient descent algorithms rely on gradient estimates whose variance decreases as fast as the objective function value. Here, we provide a unified proof of convergence for our variants of SVRG, SAGA, and MISO, and we show how to make them robust to stochastic perturbations. Specifically, we consider the minimization of a finite sum of functions as in (3), but, as explained in the previous section, each observation of the gradient $\nabla f_i(x)$ is corrupted by a random noise variable. The next proposition is inspired by the proof of SVRG [Xiao and Zhang, 2014] and characterizes the variance of g_k.

Proposition 2 (Generic variance reduction with non-uniform sampling).
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 L_i-smooth with $L_i \geq \mu$. Then, the gradient estimates g_k of the random-SVRG and MISO/SAGA/SDCA strategies defined in Section 2.3 satisfy

$$\mathbb{E}[\|g_k - \nabla f(x_{k-1})\|^2] \leq 4L_Q\mathbb{E}[F(x_{k-1}) - F^*] + \frac{2}{n}\mathbb{E}\left[\sum_{i=1}^{n}\frac{1}{nq_i}\|u_{k-1}^i - u_*^i\|^2\right] + 3\rho_Q\tilde{\sigma}^2, \tag{23}$$

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

$$\begin{aligned} u_k^i &= \nabla f_i(\tilde{x}_k) \quad \textit{for random-SVRG} \\ u_k^{j_k} &= \nabla f_{j_k}(x_k) - \beta x_k \quad \textit{and} \quad u_k^j = u_{k-1}^j \quad \textit{if} \quad j \neq j_k \quad \textit{for SAGA/MISO/SDCA}, \end{aligned}$$

and $u_*^i = \nabla f_i(x^*) - \beta x^*$ (with $\beta = 0$ for random-SVRG).

In particular, choosing the uniform distribution $q_i = 1/n$ gives $L_Q = \max_i L_i$; choosing $q_i = L_i / \sum_j L_j$ gives $L_Q = \frac{1}{n}\sum_i L_i$, which may be significantly smaller than the maximum Lipschitz constant. We note that non-uniform sampling can significantly improve the dependency of the bound to the Lipschitz constants since the average $\frac{1}{n}\sum_i L_i$ may be significantly smaller than the maximum $\max_i L_i$, but it may worsen the dependency with the variance $\tilde{\sigma}^2$ since $\rho_Q > 1$ unless Q is the uniform distribution. The proof of the proposition is given in the appendix. Next, we apply this result to Proposition 1.

Proposition 3 (Lyapunov function for variance-reduced algorithms)**.** *Consider the same setting as Proposition 2. For either variant (A) or (B) with the random-SVRG or SAGA/SDCA/MISO gradient estimators defined in Section 2.3, when using the construction of d_k from Sections 2.1 or 2.2, respectively, 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 $k \geq 1$,*

$$\frac{\delta_k}{6}\mathbb{E}[F(x_k) - F^*] + T_k \leq (1-\tau_k)\, T_{k-1} + 3\rho_Q \eta_k \delta_k \tilde{\sigma}^2 \quad \text{with} \quad \tau_k = \min\left(\delta_k, \frac{1}{5n}\right), \tag{24}$$

and

$$T_k = 5L_Q\eta_k\delta_k\mathbb{E}[F(x_k) - F^*] + \mathbb{E}[d_k(x^*) - d_k^*] + \frac{5\eta_k\delta_k}{2}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \frac{1}{q_i n}\|u_k^i - u_*^i\|^2\right].$$

The proof of the previous proposition is given in the appendix. From the Lyapunov function, we obtain a general convergence result for the variance-reduced stochastic algorithms.

Theorem 2 (Convergence of variance-reduced algorithms)**.** *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 13 to Proposition 3, by noting that for a fixed number of iterations K, the relation $\frac{\tau_k\delta_K}{6\tau_K}\mathbb{E}[F(x_k)-F^*]+T_k \leq (1-\tau_k)\,T_{k-1}+3\rho_Q\eta_k\delta_k\tilde{\sigma}^2$ is satisfied for all $k \leq K$. Indeed, $\delta_k = \frac{\tau_k\delta_k}{\tau_k} \geq \frac{\tau_k\delta_K}{\tau_K}$ since the ratio δ_t/τ_t is non-increasing. From this generic convergence theorem, we now study particular cases.

Corollary 5 (Variance-reduction, $\mu > 0$, constant step size independant of μ)**.**
Consider the same setting as in Theorem 2, where f is μ-strongly convex, $\gamma_0 = \mu$, and $\eta_k = \frac{1}{12L_Q}$. Then, with Algorithm (A),

$$\begin{aligned}\mathbb{E}\left[F(\hat{x}_k) - F^* + \alpha\|x_k - x^*\|^2\right] &\leq \Theta_k\left((1+5\tau)(F(x_0)-F^*) + \alpha\|x_0-x^*\|^2\right) + \frac{3\rho_Q\tilde{\sigma}^2}{2L_Q} \\ &\leq 8\Theta_k\left(F(x_0)-F^*\right) + \frac{3\rho_Q\tilde{\sigma}^2}{2L_Q},\end{aligned} \tag{25}$$

with $\tau = \min\left(\frac{\mu}{12L_Q}, \frac{1}{5n}\right)$, $\Theta_k = (1-\tau)^k$, and $\alpha = 36L_Q\tau \leq 3\mu$.

The proof is given in the appendix. This first corollary shows that the algorithm achieves a linear convergence rate to a noise-dominated region and produces iterates $(x_k)_{k\geq 0}$ that do not require to know the strong convexity constant μ. This shows that all estimators we consider can become *adaptive* to μ. Note that the non-uniform strategy slightly degrades the dependency in $\tilde{\sigma}^2$: indeed, $L_Q/\rho_Q = \max_{i=1} L_i$ if Q is uniform, but if $q_i = \max_i L_i / \sum_j L_j$, we have instead $L_Q/\rho_Q = \min_{i=1} L_i$. The next corollary show that a slightly better noise dependency can be achieved when the step sizes rely on μ.

Corollary 6 (Variance-reduction, $\mu > 0$, constant step size depending of μ)**.**
Consider the same setting as in Theorem 2, where f is μ-strongly convex, $\gamma_0 = \mu$, and $\eta_k = \eta = \min\left(\frac{1}{12L_Q}, \frac{1}{5\mu n}\right)$. Then, with Algorithm (A),

$$\begin{aligned}\mathbb{E}\left[F(\hat{x}_k) - F^* + 3\mu\|x_k - x^*\|^2\right] &\leq \Theta_k\left((1+5\mu\eta)(F(x_0)-F^*) + 3\mu\|x_0 - x^*\|^2\right) + 18\rho_Q\eta\tilde{\sigma}^2 \\ &\leq 8\Theta_k\left(F(x_0) - F^*\right) + 18\rho_Q\eta\tilde{\sigma}^2.\end{aligned} \tag{26}$$

We are now in shape to study a converging algorithm, with decreasing step sizes.

Corollary 7 (Variance-reduction, $\mu > 0$, decreasing step sizes)**.** *Consider the same setting as in Theorem 2, where f is μ-strongly convex 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, we use Algorithm (A) with $\gamma_0 = \mu$ and a constant step-size strategy $\eta_k = \eta$, such that the convergence rate (26) applies. Stop the optimization when we can find a point $\hat{x}_k$ such that $\mathbb{E}[F(\hat{x}_k) - F^*] \leq 24\rho_Q\eta\tilde{\sigma}^2$. Then, we restart the optimization procedure with decreasing step-sizes $\eta_k = \min\left(\eta, \frac{2}{\mu(k+2)}\right)$. The resulting number of gradient evaluations to achieve $\mathbb{E}[F(\hat{x}_k) - F^*] \leq \varepsilon$ is upper bounded by*

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

The proof is given in the appendix and shows that variance-reduction algorithms may exhibit an optimal dependency on the noise level $\tilde{\sigma}^2$ when the objective is strongly convex.

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 and 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 some parameter $\gamma_0 \geq \mu$, we consider the sequences $(\delta_k)_{k\geq 0}$ and $(\gamma_k)_{k\geq 0}$ that satisfy

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

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

$$\boxed{\begin{aligned}x_k &= \mathrm{Prox}_{\eta_k\psi}\left[y_{k-1} - \eta_k g_k\right] \quad \text{with} \quad \mathbb{E}[g_k|\mathcal{F}_{k-1}] = \nabla f(y_{k-1}) \\ y_k &= x_k + \beta_k(x_k - x_{k-1}) \quad \text{with} \quad \beta_k = \frac{\delta_k(1-\delta_k)\eta_{k+1}}{\eta_k\delta_{k+1} + \eta_{k+1}\delta_k^2},\end{aligned}} \tag{C}$$

where with constant step size $\eta_k = 1/L$, we recover a classical extrapolation parameter of accelerated gradient based methods [Nesterov, 2004]. Traditionally, estimate sequences are used to analyze the convergence of accelerated algorithms. We show in this section how to proceed for stochastic composite optimization and later, we show how to directly accelerate the random-SVRG approach we have introduced. Note that Algorithm (C) resembles the approaches introduced by Hu et al. [2009], Ghadimi and Lan [2012] but is slightly simpler since our approach involves a single extrapolation step.

4.1 Convergence analysis without variance reduction

Consider then the stochastic estimate sequence for $k \geq 1$

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

with d_0 defined as in (8) and

$$l_k(x) = f(y_{k-1}) + g_k^\top (x - y_{k-1}) + \frac{\mu}{2}\|x - y_{k-1}\|^2 + \psi(x_k) + \psi'(x_k)^\top (x - x_k), \tag{27}$$

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 (11) 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 of the estimate sequences d_k, which also appears in [Nesterov, 2004, p. 78] for constant step sizes η_k. The proof is given in the appendix.

Lemma 2 (Relations between y_k, x_k and d_k)**.** *The sequences $(x_k)_{k\geq 0}$ and $(y_k)_{k\geq 0}$ produced by Algorithm (C) satisfy for all $k \geq 0$, with $v_0 = y_0 = x_0$,*

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

Then, the next lemma is key to prove the convergence of Algorithm (C). Its proof is given in the appendix.

Lemma 3 (Key lemma for stochastic estimate sequences with acceleration)**.**
Assuming $(x_k)_{k\geq 0}$ and $(y_k)_{k\geq 0}$ are given by Algorithm (C). 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 the accelerated stochastic optimization algorithm)**.** *Under the assumptions of Lemma 2, we have for all $k \geq 1$,*

$$\mathbb{E}\left[F(x_k) - F^* + \frac{\gamma_k}{2}\|v_k - x^*\|^2\right] \leq \Gamma_k \left(F(x_0) - F^* + \frac{\gamma_0}{2}\|x_0 - x^*\|^2 + \sum_{t=1}^{k} \frac{\eta_t \sigma_t^2}{\Gamma_t}\right), \tag{28}$$

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

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

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

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

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

By Lemma 2, 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 3, 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 (28) when noticing that $\xi_k = \Gamma_k\sum_{t=1}^k \frac{\eta_t\sigma_t^2}{\Gamma_t}$. □

We now specialize the theorem to various practical cases. When not trivial, the proofs of these corollaries are given in the appendix.

Corollary 8 (Proximal accelerated SGD with constant step-size, $\mu > 0$). *Assume that f is μ-strongly convex, that the gradient estimates have constant variance $\sigma_k = \sigma$, and choose $\gamma_0 = \mu$ and $\eta_k = 1/L$ with Algorithm (C). Then,*

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k\left(F(x_0) - F^* + \frac{\mu}{2}\|x_0 - x^*\|^2\right) + \frac{\sigma^2}{\sqrt{\mu L}}. \tag{29}$$

We now show that with decreasing step sizes, we obtain an algorithm with optimal complexity similar to [Ghadimi and Lan, 2013].

Corollary 9 (**Proximal accelerated SGD with decreasing step-sizes and $\mu > 0$**). *Assume that f is μ-strongly convex and that we 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 (C), leading to the convergence rate (29), 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 resulting number of gradient evaluations to achieve $\mathbb{E}[F(x_k) - F^*] \leq \varepsilon$ is upper bounded by*

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

We note that despite the "optimal" theoretical complexity, we have observed that Algorithm (C) with the parameters of Corollaries 8 and 9 could be relatively unstable in many practical cases, as shown in Section 5, due to the large radius $\sigma^2/\sqrt{\mu L}$ of the noise region. When μ is small, such a quantity may be indeed arbitrarily larger than $F(x_0) - F^*$. Instead, we have found a minibatch strategy to be more effective in practice. When using a minibatch of size $b = \lceil L/\mu \rceil$, the theoretical complexity becomes the same as SGD, given in Corollary 2, but the algorithm enjoys the benefits of easy parallelization.

Corollary 10 (**Proximal accelerated SGD with with $\mu = 0$**). *Assume that the gradient estimates have constant variance $\sigma_k = \sigma$ and that we have a budget of K iterations for Algorithm (C). Assume that x_0 is build from an initial point $\bar{x}_0$ as in Lemma 1 with the following step size*

$$\eta_k = \eta_0 = \min\left(\frac{1}{L}, 2\sqrt{\frac{T_0}{\sigma^2}}\frac{1}{(K+1)^{3/2}}\right) \quad \text{with} \quad T_0 = \frac{1}{2}\|\bar{x}_0 - x^*\|^2.$$

Then, by choosing $\gamma_0 = \frac{1}{\eta_0}$,

$$\mathbb{E}[F(x_K) - F^*] \leq \frac{4LT_0}{(K+1)^2} + 4\sigma\sqrt{\frac{T_0}{K+1}}. \tag{30}$$

The previous convergence results are relatively similar to those obtained in [Ghadimi and Lan, 2013] for a different algorithm and is optimal for convex functions.

4.2 Accelerated algorithm with variance reduction

In this section, we show how to combine the previous methodology with variance reduction, and introduce Algorithm 3 based on random-SVRG. Then, we present the convergence analysis, which requires controlling the variance of the estimator in a similar manner to [Allen-Zhu, 2017], as stated in the next proposition. Note that the estimator does not require storing the seed of the random perturbations, unlike in the previous section.

Proposition 4 (Variance reduction for random-SVRG estimator). *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 $L_i \geq \mu$ and f is μ-strongly convex. Then, the variance of g_k defined in Algorithm 3 satisfies*

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

The proof is given in the appendix. Then, we extend Lemma 3 that was used in the previous analysis to the variance-reduction setting.

Lemma 4 (Lemma for accelerated variance-reduced stochastic optimization).
Consider the iterates provided by Algorithm 3 and call $a_k = 2L_Q\eta_k$. Then,

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

Algorithm 3 Accelerated algorithm with random-SVRG estimator

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

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

    Choose

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

    Sample i_k according to the distribution $Q = \{q_1, \ldots, q_n\}$;
    Compute the gradient estimator, possibly corrupted by stochastic perturbations:

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

    Obtain the new iterate

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

    Find the minimizer v_k of the estimate sequence d_k:

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

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

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

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

With this lemma in hand, we may now state our main convergence result.

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

$$\mathbb{E}\left[F(x_k) - F^* + \frac{\gamma_k}{2}\|v_k - x^*\|^2\right] \leq \Gamma_k \left(F(x_0) - F^* + \frac{\gamma_0}{2}\|x_0 - x^*\|^2 + \frac{3\rho_Q \tilde{\sigma}^2}{n}\sum_{t=1}^{k}\frac{\eta_t}{\Gamma_t}\right). \tag{31}$$

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,

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

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,

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

We may now use Lemma 4, which gives us

$$\begin{aligned}
\mathbb{E}[d_k^*] &\geq \left(1 - \frac{1}{n}\right)\mathbb{E}[F(\tilde{x}_{k-1})] + \frac{1}{n}\mathbb{E}[F(x_k)] + \left(\frac{1}{n}\left(\eta_k - \frac{L\eta_k^2}{2}\right) - \frac{\delta_k^2}{2\gamma_k}\right)\mathbb{E}[\|\tilde{g}_k\|^2] \\
&\qquad + \mathbb{E}\left[\tilde{g}_k^\top\left(\frac{\delta_k(1-\delta_k)\gamma_{k-1}}{\gamma_k}(v_{k-1} - y_{k-1}) + \left(\frac{1}{n} - \delta_k\right)(\tilde{x}_{k-1} - y_{k-1})\right)\right] - \xi_k,
\end{aligned} \tag{32}$$

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 (32) 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 (32) 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 (31). □

We may now derive convergence rates of our accelerated SVRG algorithm under various settings. The proofs of the following corollaries, when not straightforward, are given in the appendix. The first corollary simply uses Lemma 10.

Corollary 11 (Accelerated proximal SVRG - constant step size - $\mu > 0$)**.**
With $\eta_k = \min\left(\frac{1}{3L_Q}, \frac{1}{15\mu n}\right)$ and $\gamma_0 = \mu$, the iterates produced by Algorithm 3 satisfy

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

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \left(1 - \sqrt{\frac{5\mu}{9L_Q n}}\right)^k \left(F(x_0) - F^* + \frac{\mu}{2}\|x_0 - x^*\|^2\right) + \frac{3\rho_Q \tilde{\sigma}^2}{\sqrt{5\mu L_Q n}};$$

- *otherwise,*

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \left(1 - \frac{1}{3n}\right)^k \left(F(x_0) - F^* + \frac{\mu}{2}\|x_0 - x^*\|^2\right) + \frac{3\rho_Q \tilde{\sigma}^2}{5\mu n}.$$

The corollary uses the fact that $\Gamma_k \sum_{t=1}^k \eta/\Gamma_t \leq \eta/\delta = \sqrt{3n\eta/5\mu}$ 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 the distribution Q is uniform. When $\tilde{\sigma}^2 = 0$, the corresponding algorithm achieves the optimal complexity for finite sums [Arjevani and Shamir, 2016]. Interestingly, we see that here non-uniform sampling may hurt the convergence guarantees in some situations. Whenever $\frac{1}{\max_i L_i} > \frac{1}{5\mu n}$, the optimal sampling strategy is indeed the uniform one. Next, we show how to obtain a converging algorithm in the next corollary.

Corollary 12 (Accelerated proximal SVRG - diminishing step sizes - $\mu > 0$)**.**
Assume that f is μ-strongly convex and that we 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 3, leading to the convergence rate of the previous corollary, until $\mathbb{E}[F(x_k) - F^*] \leq B$. Then, we restart the optimization procedure with decreasing step-sizes $\eta_k = \min\left(\eta, \frac{12n}{5\mu(k+2)^2}\right)$. The resulting number of gradient evaluations to achieve $\mathbb{E}[F(x_k) - F^*] \leq \varepsilon$ is upper bounded by

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

Next, we study the case when the objective is convex, but not strongly convex.

Corollary 13 (Accelerated proximal SVRG with $\mu = 0$ and $\tilde{\sigma}^2 = 0$)**.**
Assuming $\tilde{\sigma}^2 = 0$ and that x_0 is obtained from an initial point $\bar{x}_0$ as in Lemma 1, with $\eta_0 = \frac{1}{3L_Q}$. Then, the iterates produced by Algorithm 3 with $\gamma_0 = \frac{1}{\eta_0}$ and $\eta_k = \min\left(\eta_0, \frac{1}{5\gamma_k n}\right)$ satisfy $\mathbb{E}[F(x_k) - F^*] \leq \varepsilon$ in K iterations with

$$K = O\left(n\log\left(\frac{L_Q\|\bar{x}_0 - x^*\|^2}{\varepsilon}\right) + \sqrt{\frac{nL_Q\|\bar{x}_0 - x^*\|^2}{\varepsilon}}\right). \tag{33}$$

The complexity is similar to that of Katyusha [Allen-Zhu, 2017].

Corollary 14 (Accelerated proximal SVRG - $\mu = 0$)**.**
Assuming that x_0 is obtained from an initial point $\bar{x}_0$ as in Lemma 1, with $\eta_0 \leq \frac{1}{3L_Q}$ specified below, and $\sigma_0^2 = \tilde{\sigma}^2/n$ (which can be achieved by computing n individual gradients). We introduce the quantity $K_0 = 6n\log(5n)$ and assume that one has a budget of $K \geq K_0$ iterations for Algorithm 3. Choose then a constant step size policy for $k \leq K$:

$$\eta_k = \eta_0 = \min\left(\frac{1}{3L_Q}, \frac{n\sqrt{6T_0}}{\sqrt{\rho_Q \sigma^2}(K+1)^{3/2}}\right) \quad \textit{with} \quad T_0 = \frac{1}{2}\|\bar{x}_0 - x^*\|^2. \tag{34}$$

Then,

$$\mathbb{E}[F(x_K) - F^*] \leq \frac{18nL_Q}{(K+1)^2}T_0 + \frac{\tilde{\sigma}\sqrt{24\rho_Q T_0}}{\sqrt{K+1}}. \tag{35}$$

As earlier when studying the stochastic setting with $\mu = 0$, the theoretical analysis is not a practical one, but simply illustrates the dependency we achieve with respect to $\tilde{\sigma}$ when using a constant step size strategy.

5 Experiments

In this section, we evaluate numerically the approaches introduced in the previous sections.

5.1 Datasets, Formulations, and Methods

Following classical benchmarks in optimization methods for machine learning [see, *e.g.* Schmidt et al., 2017], we consider a empirical risk minimization formulations. Given training data $(a_i, b_i)_{i=1,\dots,n}$, with a_i in $\mathbb{R}^p$ and b_i in $\{-1, +1\}$, we consider the optimization problem

$$\min_{x\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \phi(b_i a_i^\top x) + \frac{\lambda}{2}\|x\|^2,$$

where ϕ is either the logistic loss $\phi(u) = \log(1 + e^{-u})$, or the squared hinge loss $\phi(u) = \max(0, 1-u)^2$. Both functions are L-smooth; when the vectors a_i have unit norm, we may indeed choose $L = 0.25$ for the logistic loss and $L = 1$ for the squared hinge loss. Using the squared hinge loss in addition to the logistic one is interesting; whereas the logistic loss has bounded gradients on $\mathbb{R}^p$, this is not the case for the squared hinge loss. With unbounded optimization domain, the gradient norms may be indeed large in some regions of the solution space, which may lead in turn to large variance σ^2 of the gradient estimates obtained SGD, causing instabilities.

The scalar λ is a regularization parameter that acts as a lower bound on the strong convexity constant of the problem. We consider the parameters $\mu = \lambda = 1/10n$ in our problems, which is of the order of the smallest values that one would try when doing a parameter search, *e.g.*, by cross-validation. For instance, this is empirically observed for the dataset cifar-ckn described below, where a test set is available, allowing us to check that the "optimal" regularization parameter leading to the lowest generalization error is indeed of this order. We also report an experiment with $\lambda = 1/100n$ in order to study the effect of the problem conditioning on the methods performance.

Following Bietti and Mairal [2017], Zheng and Kwok [2018], we consider DropOut perturbations [Srivastava et al., 2014] to illustrate the robustness to noise of the algorithms. DropOut consists of randomly setting to zero each entry of a data point with probability δ, leading to the optimization problem

$$\min_{x\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\rho\left[\phi(b_i(\rho \circ a_i)^\top x)\right] + \frac{\lambda}{2}\|x\|^2, \tag{36}$$

where ρ is a binary vector in $\{0,1\}^p$ with i.i.d. Bernoulli entries, and $\circ$ denotes the elementwise multiplication between two vectors. We consider two DropOut regimes, with δ in $\{0.01, 0.1\}$, representing small and medium perturbations, respectively.

Then, 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[3] and contains $n = 250\,000$ points in dimension $p = 500$.
- gene is 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[4] 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 to be a good proxy of the computational time. Indeed, (i) if one is allowed to store the variables z_i^k, then n iterations exactly 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 or by using more efficient routines (*e.g.*, BLAS2 vs BLAS1 routines for linear algebra). Each experiment is conducted five times and we always report the average of the five experiments in each figure.

5.2 Evaluation of Algorithms without Perturbations

First, we study the behavior of all methods when $\tilde{\sigma}^2 = 0$. We report the corresponding results in Figures 1, 2, and 3. Since the problem is deterministic, we can check that the value F^* we consider is indeed optimal by computing a duality gap using Fenchel duality. For SGD and random-SVRG, we do not use any averaging strategy, which we found to empirically slow down convergence, when used from the start; knowing when to start averaging is indeed not easy and requires heuristics which we do not evaluate here.

From these experiments, we obtain the following conclusions:

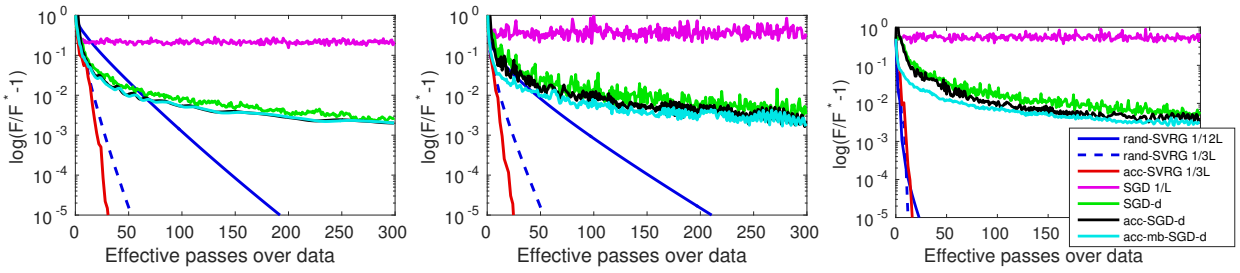
- Acceleration for SVRG is effective on the datasets gene and ckn-cifar except on alpha, where all SVRG-like methods perform already well. This may be due to strong convexity hidden in alpha leading to a regime where acceleration does not occur—that is, when the complexity is $O(n \log(1/\varepsilon))$, which is independent of the condition number.
- Acceleration is more effective when the problem is badly conditioned. When $\lambda = 1/100n$, acceleration brings several orders of magnitude improvement in complexity.
- Accelerated SGD is unstable with the squared hinge loss. During the initial phase with constant step size $1/L$, the expected primal gap is in a region of radius $O(\sigma^2/\sqrt{\mu L} \approx \sqrt{n}\sigma^2$, which is potentially huge, causing large gradients and instabilities.
- Accelerated minibatch SGD performs best among the SGD methods and is competitive with SVRG in the low precision regime.

[3] `http://largescale.ml.tu-berlin.de/`

[4] `https://www.cs.toronto.edu/~kriz/cifar.html`

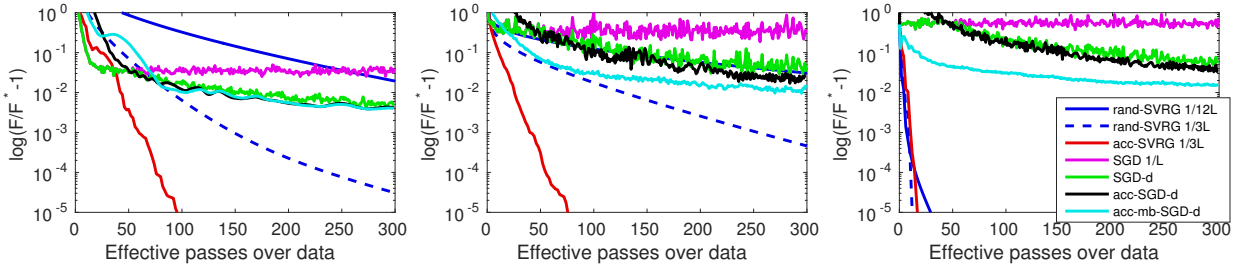
Algorithm	step size η_k	Theory	Complexity $O(.)$	Bias $O(.)$
SGD	$\frac{1}{L}$	Cor. 1	$\frac{L}{\mu}\log\left(\frac{C_0}{\varepsilon}\right)$	$\frac{\sigma^2}{L}$
SGD-d	$\min\left(\frac{1}{L}, \frac{2}{\mu(k+2)}\right)$	Cor. 2	$\frac{L}{\mu}\log\left(\frac{C_0}{\varepsilon}\right) + \frac{\sigma^2}{\mu\varepsilon}$	0
acc-SGD	$\frac{1}{L}$	Cor. 8	$\sqrt{\frac{L}{\mu}}\log\left(\frac{C_0}{\varepsilon}\right)$	$\frac{\sigma^2}{\sqrt{\mu L}}$
acc-SGD-d	$\min\left(\frac{1}{L}, \frac{4}{\mu(k+2)^2}\right)$	Cor. 9	$\sqrt{\frac{L}{\mu}}\log\left(\frac{C_0}{\varepsilon}\right) + \frac{\sigma^2}{\mu\varepsilon}$	0
acc-mb-SGD-d	$\min\left(\frac{1}{L}, \frac{4}{\mu(k+2)^2}\right)$	Cor. 9	$\frac{L}{\mu}\log\left(\frac{C_0}{\varepsilon}\right) + \frac{\sigma^2}{\mu\varepsilon}$	0
rand-SVRG	$\frac{1}{12L}$	Cor. 5	$\left(n + \frac{L}{\mu}\right)\log\left(\frac{C_0}{\varepsilon}\right)$	$\frac{\tilde{\sigma}^2}{L}$
rand-SVRG-d	$\min\left(\frac{1}{12L_Q}, \frac{1}{5\mu n}, \frac{2}{\mu(k+2)}\right)$	Cor. 7	$\left(n + \frac{L}{\mu}\right)\log\left(\frac{C_0}{\varepsilon}\right) + \frac{\tilde{\sigma}^2}{\mu\varepsilon}$	0
acc-SVRG	$\min\left(\frac{1}{3L_Q}, \frac{1}{15\mu n}\right)$	Cor. 11	$\left(n + \sqrt{\frac{nL}{\mu}}\right)\log\left(\frac{C_0}{\varepsilon}\right)$	$\frac{\tilde{\sigma}^2}{\sqrt{n\mu L}+n\mu}$
acc-SVRG-d	$\min\left(\frac{1}{3L_Q}, \frac{1}{15\mu n}, \frac{12n}{5\mu(k+2)^2}\right)$	Cor. 12	$\left(n + \sqrt{\frac{nL}{\mu}}\right)\log\left(\frac{C_0}{\varepsilon}\right) + \frac{\tilde{\sigma}^2}{\mu\varepsilon}$	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 $\eta = 1/3L$, even though our analysis requires $\eta \leq 1/12L$. 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 $\tilde{\sigma} \ll \sigma$.



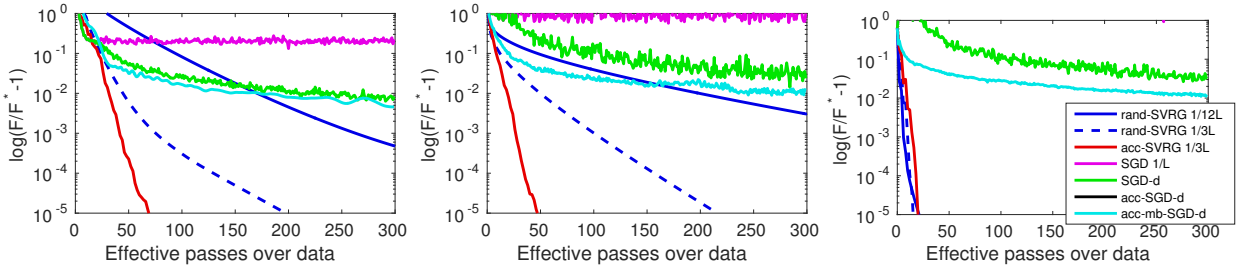
(a) Dataset gene (b) Dataset ckn-cifar (c) Dataset alpha

Figure 1: Optimization curves without perturbations when using the logistic loss and the parameter $\lambda = 1/10n$. We plot the value of the objective function on a logarithmic scale as a function of the effective passes over the data (see main text for details). Best seen in color by zooming on a computer screen.



(a) Dataset gene (b) Dataset ckn-cifar (c) Dataset alpha

Figure 2: Same experiment as in Figure 1 with $\lambda = 1/100n$.

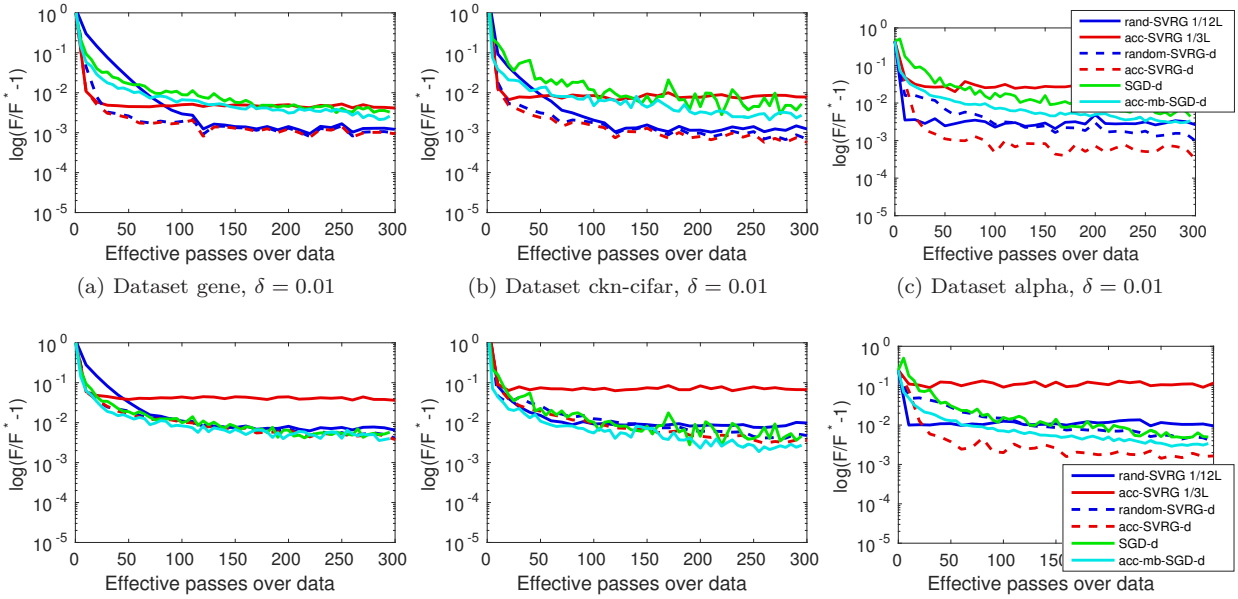


(a) Dataset gene (b) Dataset ckn-cifar (c) Dataset alpha

Figure 3: Same experiment as in Figure 1 with squared hinge loss instead of logistic. acc-SGD-d was unstable for this setting due to the large size of the noise region $\sigma^2/\sqrt{\mu L} = \sqrt{10n}\sigma^2$ and potentially large gradients of the loss function over the optimization domain.

5.3 Evaluation of Algorithms with Perturbations

We now consider the same setting as in the previous section, but we add DropOut perturbations with rate δ in $\{0.01, 0.1\}$. As predicted by theory, all approaches with constant step size do not converge. Therefore, we only report the results for rand-SVRG and acc-SVRG in such a constant step size regime. Then, we consider the different algorithms with decreasing step sizes and report the results in Figures 4, 5, and 6. We evaluate the loss function every 5 data passes and we estimate the expectation (36) 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^*.

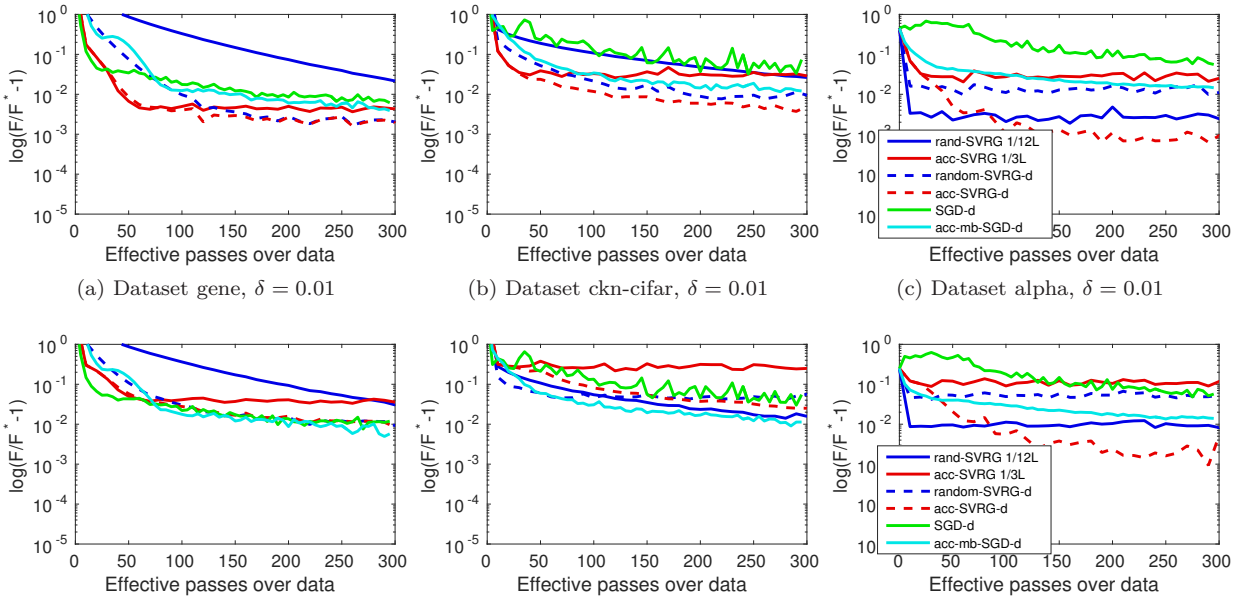


(a) Dataset gene, $\delta = 0.01$ (b) Dataset ckn-cifar, $\delta = 0.01$ (c) Dataset alpha, $\delta = 0.01$

(d) Dataset gene, $\delta = 0.1$ (e) Dataset ckn-cifar, $\delta = 0.1$ (f) Dataset alpha, $\delta = 0.1$

Figure 4: Optimization curves with DropOut rate δ when using the logistic loss and $\lambda = 1/10n$. We plot the value of the objective function on a logarithmic scale as a function of the effective passes over the data. Best seen in color by zooming on a computer screen.

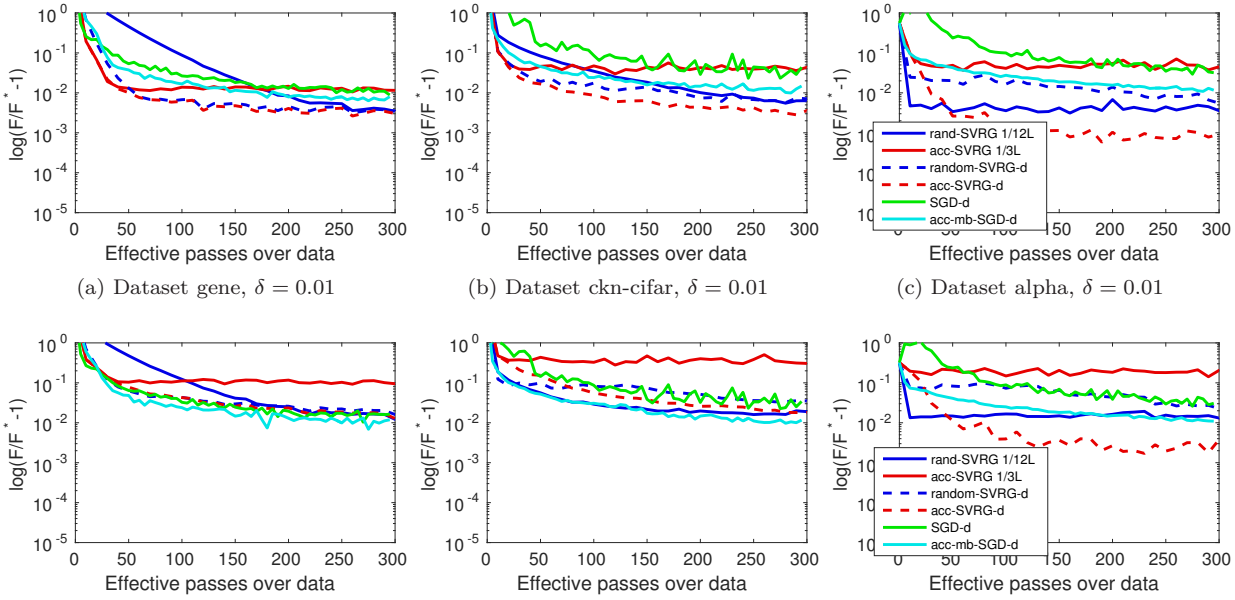
The conclusions of these experiments are the following:



(a) Dataset gene, $\delta = 0.01$ (b) Dataset ckn-cifar, $\delta = 0.01$ (c) Dataset alpha, $\delta = 0.01$

(d) Dataset gene, $\delta = 0.1$ (e) Dataset ckn-cifar, $\delta = 0.1$ (f) Dataset alpha, $\delta = 0.1$

Figure 5: Same setting as in Figure 4 but with $\lambda = 1/100n$.



(a) Dataset gene, $\delta = 0.01$ (b) Dataset ckn-cifar, $\delta = 0.01$ (c) Dataset alpha, $\delta = 0.01$

(d) Dataset gene, $\delta = 0.1$ (e) Dataset ckn-cifar, $\delta = 0.1$ (f) Dataset alpha, $\delta = 0.1$

Figure 6: Same setting as in Figure 4 but with the squared hinge loss.

- accelerated minibatch SGD performs the best among SGD approaches in general.
- accelerated SVRG performs better than SVRG in general, or they achieve the same performance. As in the deterministic case, the gains are typically more important in ill-conditioned cases.
- accelerated SVRG performs uniformly better than SGD approaches in the low perturbation regime $\delta = 0.01$ and only on the alpha dataset when $\delta = 0.1$. Otherwise, the methods perform similarly.
- not reported on these figures, high perturbation regimes, *e.g.*, $\delta = 0.3$ make variance reduction less useful since the noise due to data sampling becomes potentially of the same order as $\tilde{\sigma}^2$; Yet, benefits are still seen on the alpha dataset, whereas SGD approaches perform slightly better than SVRG approaches on ckn-cifar and gene.

6 Discussion

In this paper, we have studied simple stochastic gradient-based rules with or without variance reduction, and presented an accelerated algorithm dedicated to finite-sums minimization under the presence of stochastic perturbations. The approach we propose achieves the classical optimal worst-case complexities for finite-sum optimization when there is no perturbation [Arjevani and Shamir, 2016], and exhibit an optimal dependency in the noise variance $\tilde{\sigma}^2$ for convex and strongly convex problems.

Our work is based on stochastic variants of estimate sequences introduced by Nesterov [1983, 2004]. The framework leads naturally to many algorithms with relatively generic proofs of convergence, where convergence is proven at the same time as the algorithm's design. With simple iterate averaging techniques inspired by Ghadimi and Lan [2013], we show that a large class of variance-reduction stochastic optimization methods can be made robust to stochastic perturbations. Estimate sequences also naturally lead to several accelerated algorithms, some of them we did not present in this paper. For instance, it is possible to show that replacing in (27) the lower bound $\psi(x_k) + \psi'(x_k)^\top(x - x_k)$ by $\psi(x)$ itself—in a similar way as we proceeded to obtain iteration (B) from iteration (A)—also leads to an accelerated algorithm with similar guarantees as (C).

Possibilities offered by estimate sequences are large, but our framework also admits a few limitations, paving the way for future work. In particular, our results are currently limited to Euclidean metrics—meaning that our convergence rates typically depend on quantities involving the Euclidean norm (e.g., strong convexity or L-smooth inequalities), and one may expect extensions of our work to other metrics such as Bregman distances. Estimate sequences admit indeed known extensions to such metrics, and can also deal with higher-order smoothness assumptions than Lipschitz continuity of the gradient [Baes, 2009]—*e.g.*, cubic regularization [Nesterov and Polyak, 2006]. We leave such directions for the future.

Another limitation we encountered was the inability to propose robust accelerated variants of SAGA, MISO, or SDCA based on our framework. We believe this task to be possible and we are planning to explore such a direction by making connections with the Catalyst framework of Lin et al. [2018], which achieves generic acceleration for deterministic problems, at the price of a logarithmic factor in the optimal complexity.

Acknowledgments

This work was supported by the ERC grant SOLARIS (number 714381) and a grant from ANR (MACARON project ANR-14-CE23-0003-01). The authors would like to thank Anatoli Juditsky for numerous interesting discussions that greatly improved the quality of this manuscript.

References

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

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

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

M. Baes. Estimate sequence methods: extensions and approximations. *ETH technical report*, 2009.

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

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

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

D. Csiba, Z. Qu, and P. Richtárik. Stochastic dual coordinate ascent with adaptive probabilities. In *International Conference on Machine Learning (ICML)*, 2015.

A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, 2014a.

A. Defazio, T. Caetano, and J. Domke. Finito: A faster, permutable incremental gradient method for big data problems. In *Proceedings of the International Conferences on Machine Learning (ICML)*, 2014b.

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

S. Ghadimi and G. Lan. 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.

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

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

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

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

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

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

H. Lin, J. Mairal, and Z. Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research (JMLR)*, 18(212):1–54, 2018.

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

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

J. Mairal, F. Bach, and J. Ponce. Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

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

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

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

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

Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

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

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

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

A. Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

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

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

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

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

V. Vapnik. *The nature of statistical learning theory*. Springer, 2000.

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

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

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

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

A Useful Mathematical Results

A.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 5 (**Quadratic upper bound for L-smooth functions).**
Let $f : \mathbb{R}^p \to \mathbb{R}$ be L-smooth. Then, for all x, x' in $\mathbb{R}^p$,

$$|f(x') - f(x) - \nabla f(x)^\top (x' - x)| \leq \frac{L}{2}\|x - x'\|_2^2.$$

Lemma 6 (**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 $\mathbb{R}^p$:

$$f(x') \geq f(x) + z^\top (x' - x) + \frac{\mu}{2}\|x - x'\|_2^2.$$

Lemma 7 (**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 $\mathcal{X}$:

$$f(x) \geq f(x^*) + \frac{\mu}{2}\|x - x^*\|_2^2.$$

Lemma 8 (**Useful inequality for smooth and convex functions).**
Consider an L-smooth μ-strongly convex function f defined on $\mathbb{R}^p$ and a parameter β in $[0, \mu]$. Then, for all x, y in $\mathbb{R}^p$,

$$\|\nabla f(x) - \nabla f(y) - \beta(x - y)\|^2 \leq 2L(f(x) - f(y) - \nabla f(y)^\top (x - y)).$$

Proof. Let us define the function $\phi(x) = f(x) - \frac{\beta}{2}\|x\|^2$, which is $(\mu - \beta)$-strongly convex. It is then easy to show that ϕ is $(L - \beta)$-smooth, according to Theorem 2.1.5 in [Nesterov, 2004]: indeed, for all x, y in $\mathbb{R}^p$,

$$\begin{aligned}\phi(x) = f(x) - \frac{\beta}{2}\|x\|^2 &\leq f(y) + \nabla f(y)^\top (x - y) + \frac{L}{2}\|x - y\|^2 - \frac{\beta}{2}\|x\|^2 \\ &= \phi(y) + \nabla \phi(y)^\top (x - y) + \frac{L - \beta}{2}\|x - y\|^2,\end{aligned}$$

and again according to Theorem 2.1.5 of [Nesterov, 2004],

$$\begin{aligned}\|\nabla \phi(x) - \nabla \phi(y)\|^2 &\leq 2L(\phi(x) - \phi(y) - \nabla \phi(y)^\top (x - y)) \\ &= 2L\left(f(x) - f(y) - \nabla f(y)^\top (x - y) - \frac{\beta}{2}\|x - y\|^2\right) \\ &\leq 2L\left(f(x) - f(y) - \nabla f(y)^\top (x - y)\right).\end{aligned}$$

□

A.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 9 (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 10 (Simple relation)**.** *Consider a sequence of weights $(\delta_k)_{k\geq 0}$ in $(0,1)$. Then,*

$$\sum_{t=1}^{k}\frac{\delta_t}{\Gamma_t} + 1 = \frac{1}{\Gamma_k} \qquad \text{where} \qquad \Gamma_t := \prod_{i=1}^{t}(1-\delta_i). \tag{37}$$

Lemma 11 (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 $k \geq 0$,*

$$\Gamma_k \leq \min\left((1-\mu\eta)^k, \frac{1}{1+\gamma_0\eta k}\right). \tag{38}$$

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 (38). 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 12 (Accelerated convergence rate of Γ_k)**.** *Consider the same quantities defined in Lemma 10 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]. □

A.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 13 (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 $k \geq 1$,*

$$\frac{\delta_k}{\alpha}\mathbb{E}[F(x_k) - F^*] + T_k \leq (1-\delta_k)T_{k-1} + \beta_k, \tag{39}$$

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

No averaging.

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

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,

$$\mathbb{E}[F(\hat{x}_k) - F^*] + \alpha T_k \leq \Gamma_k\left(\alpha T_0 + \mathbb{E}[F(x_0) - F^*] + \alpha\sum_{t=1}^{k}\frac{\beta_t}{\Gamma_t}\right). \tag{40}$$

Proof. Given that $T_k \leq (1-\delta_k)T_{k-1} + \beta_k$, we obtain (39) by simply unrolling the recursion. To analyze the effect of the averaging strategies, divide now (39) 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 (37), we may then use Jensen's inequality and we obtain (40). ∎

B Proofs of the Main Results

B.1 Proof of Lemma 1

Proof.

$$\begin{aligned}
\mathbb{E}[F(x_0)] &= \mathbb{E}[f(x_0) + \psi(x_0)] \\
&\leq \mathbb{E}\left[f(\bar{x}_0) + \nabla f(\bar{x}_0)^\top (x_0 - \bar{x}_0) + \frac{L}{2}\|x_0 - \bar{x}_0\|^2 + \psi(x_0)\right] \\
&= \mathbb{E}\left[f(\bar{x}_0) + g_0^\top (x_0 - \bar{x}_0) + \frac{L}{2}\|x_0 - \bar{x}_0\|^2 + \psi(x_0)\right] + \mathbb{E}\left[(\nabla f(\bar{x}_0) - g_0)^\top (x_0 - \bar{x}_0)\right] \\
&= \mathbb{E}\left[f(\bar{x}_0) + g_0^\top (x_0 - \bar{x}_0) + \frac{L}{2}\|x_0 - \bar{x}_0\|^2 + \psi(x_0)\right] + \mathbb{E}\left[(\nabla f(\bar{x}_0) - g_0)^\top x_0\right] \\
&= \mathbb{E}\left[f(\bar{x}_0) + g_0^\top (x_0 - \bar{x}_0) + \frac{L}{2}\|x_0 - \bar{x}_0\|^2 + \psi(x_0)\right] + \mathbb{E}\left[(\nabla f(\bar{x}_0) - g_0)^\top (x_0 - w_0)\right] \\
&\leq \mathbb{E}\left[f(\bar{x}_0) + g_0^\top (x_0 - \bar{x}_0) + \frac{L}{2}\|x_0 - \bar{x}_0\|^2 + \psi(x_0)\right] + \mathbb{E}\left[\|\nabla f(\bar{x}_0) - g_0\|\|x_0 - w_0\|\right] \\
&\leq \mathbb{E}\left[f(\bar{x}_0) + g_0^\top (x_0 - \bar{x}_0) + \frac{L}{2}\|x_0 - \bar{x}_0\|^2 + \psi(x_0)\right] + \mathbb{E}\left[\eta_0\|\nabla f(\bar{x}_0) - g_0\|^2\right] \\
&\leq \mathbb{E}\left[f(\bar{x}_0) + g_0^\top (x_0 - \bar{x}_0) + \frac{1}{2\eta_0}\|x_0 - \bar{x}_0\|^2 + \psi(x_0)\right] + \eta_0\sigma_0^2,
\end{aligned}$$

where $w_0 = \mathrm{Prox}_{\eta_0\psi}[\bar{x}_0 - \eta_0\nabla f(\bar{x}_0)]$. The first inequality is due to the L-smoothness of f (Lemma 5); then, the next three relations exploit the fact that $\mathbb{E}[(\nabla f(\bar{x}_0) - g_0)^\top z = 0$ for all z that is deterministic (which is the case for $\bar{x}_0$ and w_0); the third inequality uses the non-expansiveness of the proximal operator. Then, note that x_0 minimizes the strongly convex function $x \mapsto g_0^\top (x - \bar{x}_0) + \frac{1}{2\eta_0}\|x - \bar{x}_0\|^2 + \psi(x)$ such that, from Lemma 7,

$$\begin{aligned}
\mathbb{E}[F(x_0)] &\leq \mathbb{E}\left[f(\bar{x}_0) + g_0^\top (x^* - \bar{x}_0) + \frac{1}{2\eta_0}\|x^* - \bar{x}_0\|^2 + \psi(x^*) - \frac{1}{2\eta_0}\|x^* - x_0\|^2\right] + \eta_0\sigma_0^2 \\
&= \mathbb{E}\left[f(\bar{x}_0) + \nabla f(\bar{x}_0)^\top (x^* - \bar{x}_0) + \frac{1}{2\eta_0}\|x^* - \bar{x}_0\|^2 + \psi(x^*) - \frac{1}{2\eta_0}\|x^* - x_0\|^2\right] + \eta_0\sigma_0^2 \\
&\leq \mathbb{E}\left[f(x^*) + \frac{1}{2\eta_0}\|x^* - \bar{x}_0\|^2 + \psi(x^*) - \frac{1}{2\eta_0}\|x^* - x_0\|^2\right] + \eta_0\sigma_0^2 \\
&= F^* + \mathbb{E}\left[\frac{1}{2\eta_0}\|x^* - \bar{x}_0\|^2 - \frac{1}{2\eta_0}\|x^* - x_0\|^2\right] + \eta_0\sigma_0^2.
\end{aligned}$$

□

B.2 Proof of Corollary 2

Proof. Given the linear convergence rate (20), 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 9, 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 9 and 10. 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. □

B.3 Proof of Proposition 2

Proof. The proof borrows a large part of the analysis of Xiao and 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 in the generic form (15), with $\beta = 0$ for SAGA or SVRG. 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}) - \beta x_{k-1} - z_{k-1}^{i_k}) + \bar{z}_{k-1} + \beta x_{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}) - \beta x_{k-1} - z_{k-1}^{i_k}) + \bar{z}^{k-1} + \beta x_{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}) - \beta x_{k-1} - z_{k-1}^{i_k}) + \bar{z}^{k-1} + \beta x_{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}) - \beta 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}) - \beta 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}) - \beta x_{k-1} - u_*^i + u_*^i - z_{k-1}^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}) - \beta 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^*) - \beta(x_{k-1} - 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 Lemma 8.

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 (23). □

B.4 Proof of Proposition 3

Proof. To make the notation more compact, we call

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

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,

$$\delta_k F_k + D_k \leq (1 - \delta_k)D_{k-1} + 4L_Q\eta_k\delta_k F_{k-1} + 2\eta_k\delta_k C_{k-1} + 3\rho_Q\eta_k\delta_k\tilde{\sigma}^2. \tag{41}$$

Then, we note that both for the SVRG and SAGA/MISO/SDCA strategies, we have (with $\beta = 0$ for SVRG),

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

By taking a weighted average, this yields

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

where the second inequality comes from Lemma 8 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 (41) with some $\beta_k > 0$ that we will specify later:

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

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,

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

which immediately yields (24) with the appropriate definition of T_k, and by noting that $(1 - 10L_Q\eta_k) \geq \frac{1}{6}$. □

B.5 Proof of Corollary 5

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

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

where the inequality comes from Lemma 8 and the definition of the u_0^i's. Then, we immediately obtain the first line of (25). Then, we conclude by noting that $5\tau \leq 1$, and that $\alpha \leq 3\mu$ and we use Lemma 7. □

B.6 Proof of Corollary 6

Proof. We follow similar steps as in the proof of Corollary 5. 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

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

Then, we use the same upper-bound on T_0 as in the proof of Corollary 5, 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. □

B.7 Proof of Corollary 7

Proof. Since the convergence rate (26) 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 7, we have $6T_0 \leq 7(F(x_0) - F^*)$. Then, after taking the expectation with respect to the output of the first stage,

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

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 9, for $k \geq k_0$,

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

which gives the desired complexity.

□

B.8 Proof of Lemma 2

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

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

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

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

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

B.9 Proof of Lemma 3

Proof. Following similar steps as in the proof of Lemma 1 (since the relation between x_k and y_{k-1} is the same as the relation between x_0 and $\bar{x}_0$ in this other lemma), we have

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

where we use the fact that $x_k = y_{k-1} - \eta_k \tilde{g}_k$. □

B.10 Proof of Corollary 9

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 9. 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 9 and 10. This leads to the desired iteration complexity. □

B.11 Proof of Corollary 10

Proof. We note that according to Lemma 12, we have

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

Then, we apply Theorem 3, we obtain the relation

$$\begin{aligned}\mathbb{E}[F(x_K) - F^*] &\leq \Gamma_K \left(F(x_0) - F^* + \frac{\gamma_0}{2}\|x_0 - x^*\|^2\right) + \sigma^2\eta\Gamma_K \sum_{t=1}^{K} \frac{1}{\Gamma_t} \\ &\leq \Gamma_K \left(\frac{1}{\eta}T_0 + \eta\sigma^2\right) + \sigma^2\eta K \\ &\leq \frac{4}{(1+K)^2\eta}T_0 + \sigma^2\eta(K+1)\end{aligned}$$

Optimizing with respect to η under the constraint $\eta \leq 1/L$ gives the optimal value of $\eta = \min\left(\frac{1}{L}, 2\sqrt{\frac{T_0}{\sigma^2}}\frac{1}{(K+1)^{3/2}}\right)$. It is then easy to check that for both potential values of η, Eq. (30) is satisfied. □

B.12 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

$$\begin{aligned}\sigma_k^2 &\leq \mathbb{E}\left\|\frac{1}{q_{i_k}n}\left(\nabla f_{i_k}(y_{k-1}) - \nabla f_{i_k}(\tilde{x}_{k-1})\right)\right\|^2 + \mathbb{E}[\|\bar{\zeta}_{k-1}\|^2] + 2\rho_Q\tilde{\sigma}^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{q_i n}\mathbb{E}\left\|\nabla f_i(y_{k-1}) - \nabla f_i(\tilde{x}_{k-1})\right\|^2 + 3\rho_Q\tilde{\sigma}^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\frac{2L_i}{q_i n}\mathbb{E}\left[f_i(\tilde{x}_{k-1}) - f_i(y_{k-1}) - \nabla f_i(y_{k-1})^\top(\tilde{x}_{k-1} - y_{k-1})\right] + 3\rho_Q\tilde{\sigma}^2 \qquad (42) \\ &\leq \frac{1}{n}\sum_{i=1}^{n}2L_Q\mathbb{E}\left[f_i(\tilde{x}_{k-1}) - f_i(y_{k-1}) - \nabla f_i(y_{k-1})^\top(\tilde{x}_{k-1} - y_{k-1})\right] + 3\rho_Q\tilde{\sigma}^2 \\ &= 2L_Q\mathbb{E}\left[f(\tilde{x}_{k-1}) - f(y_{k-1}) - \nabla f(y_{k-1})^\top(\tilde{x}_{k-1} - y_{k-1})\right] + 3\rho_Q\tilde{\sigma}^2 \\ &= 2L_Q\mathbb{E}\left[f(\tilde{x}_{k-1}) - f(y_{k-1}) - g_k^\top(\tilde{x}_{k-1} - y_{k-1})\right] + 3\rho_Q\tilde{\sigma}^2,\end{aligned}$$

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]. □

B.13 Proof of Lemma 4

Proof. We can show that Lemma 3 still holds and thus,

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

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. □

B.14 Proof of Corollary 12

Proof. The proof is similar to that of Corollary 9 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 9. This yields for $k \geq k_0 = \left\lceil\sqrt{\frac{12n}{5\mu\eta}} - 2\right\rceil$,

$$\begin{aligned}\mathbb{E}[F(x_k) - F^*] &\leq \Gamma_k\left(\mathbb{E}\left[F(x_0) - F^* + \frac{\mu}{2}\|x_0 - x^*\|^2\right] + \frac{3\rho_Q\tilde{\sigma}^2}{n}\sum_{t=1}^{k}\frac{\eta_t}{\Gamma_t}\right) \\ &\leq \Gamma_k\left(2B + \frac{3\rho_Q\tilde{\sigma}^2\eta}{n}\sum_{t=1}^{k_0-1}\frac{1}{\Gamma_t} + \frac{3\rho_Q\tilde{\sigma}^2}{n}\sum_{t=k_0}^{k}\frac{12n}{5\Gamma_t\mu(t+2)^2}\right) \\ &= \frac{k_0(k_0+1)}{(k+1)(k+2)}\left(\Gamma_{k_0-1}2B + \frac{3\rho_Q\tilde{\sigma}^2\eta}{n}\Gamma_{k_0-1}\sum_{t=1}^{k_0-1}\frac{1}{\Gamma_t}\right) + \frac{36\rho_Q\tilde{\sigma}^2}{5\mu}\sum_{t=k_0}^{k}\frac{\Gamma_k}{\Gamma_t(t+2)^2} \\ &= \frac{k_0(k_0+1)}{(k+1)(k+2)}\left(\Gamma_{k_0-1}2B + (1-\Gamma_{k_0-1})\frac{3\rho_Q\tilde{\sigma}^2\eta}{n\delta_{k_0}}\right) + \frac{36\rho_Q\tilde{\sigma}^2}{5\mu}\sum_{t=k_0}^{k}\frac{\Gamma_k}{\Gamma_t(t+2)^2} \\ &\leq \frac{2k_0(k_0+1)B}{(k+1)(k+2)} + \frac{8\rho_Q\tilde{\sigma}^2}{\mu(k+1)(k+2)}\left(\sum_{t=k_0+1}^{k}\frac{(t+1)(t+2)}{(t+2)^2}\right) \\ &\leq \frac{2k_0B}{k+2} + \frac{8\rho_Q\tilde{\sigma}^2}{\mu(k+2)},\end{aligned}$$

where we use Lemmas 9 and 10. Then, note that $k_0B \leq 6\rho_Q\tilde{\sigma}^2/\mu$ and we obtain the right iteration complexity. □

B.15 Proof of Corollary 13

Proof. We first write the proof in a more general setting since we will obtain an intermediate result that will be useful for the proof of Corollary 14, before setting $\eta_0 = \frac{1}{3L_Q}$ and $\tilde{\sigma}^2 = 0$. According to Lemma 1, assuming $\sigma_0^2 = \tilde{\sigma}^2/n$, the first stochastic gradient descent step ensures that

$$\mathbb{E}\left[F(x_0) - F^* + \frac{\gamma_0}{2}\|x_0 - x^*\|^2\right] \leq \frac{1}{2\eta_0}\|\bar{x}_0 - x^*\|^2 + \frac{\eta_0\tilde{\sigma}^2}{n}.$$

The computational cost of this initialization step is n gradients. Then, it is easy to show from Theorem 4 that

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \frac{\Gamma_k}{2\eta_0}\|\bar{x}_0 - x^*\|^2 + \frac{\eta_0\rho_Q\tilde{\sigma}^2(k+1)}{n}. \tag{43}$$

Then, we have a regime where the step sizes $\eta_k = \min\left(\eta_0, \frac{1}{5\gamma_k n}\right) = \frac{1}{5\gamma_k n}$, with $\gamma_k = \gamma_0\Gamma_k$, which corresponds to $\delta_k = \frac{1}{3n}$ and a linear rate of convergence for $\Gamma_k = \left(1 - \frac{1}{3n}\right)^k$. We may now call k_0 the index when the step size switches to the constant regime $\eta_k = \eta_0$. This index is such that $\eta_0 \leq \frac{1}{5\gamma_k n} = \frac{\eta_0}{5\Gamma_{k_0} n}$ and thus it is the index such that $\left(1 - \frac{1}{3n}\right)^{k_0} \leq \frac{1}{5n} < \left(1 - \frac{1}{3n}\right)^{k_0 - 1}$, which gives us $k_0 = O(n \log n)$.

It remains then to characterize the rate of convergence of $(\Gamma_k)_{k\geq 0}$. We of course have $\Gamma_k = \left(1 - \frac{1}{3n}\right)^k$ for $k < k_0$, and when $k \geq k_0$, it is possible to use Lemma 12 to obtain

$$\Gamma_k = \Gamma_{k_0 - 1}\frac{4}{\left(2 + (k + 1 - k_0)\sqrt{\frac{5\gamma_{k_0-1}\eta}{3n}}\right)^2} \leq \Gamma_{k_0-1}\frac{4}{(k+1-k_0)^2\frac{5\Gamma_{k_0-1}}{3n}} = \frac{3n}{(k+1-k_0)^2}.$$

When $\tilde{\sigma}^2 = 0$, this gives us the complexity (33). ☐

B.16 Proof of Corollary 14

Proof. We start from the proof of Corollary 14 and in particular with (43). Still following this proof, we note that $K \geq K_0 \geq 2k_0$ by noting that $-\log(5n) \leq k_0 \log\left(1 - \frac{1}{3n}\right) \leq -\frac{k_0}{3n}$. Therefore $K - k_0 \geq \frac{K}{2}$ and $\Gamma_K \leq \frac{3n}{(K+1-k_0)^2} \leq \frac{6n}{(K+1)^2}$. Thus,

$$\mathbb{E}\left[F(x_K) - F^*\right] \leq \frac{6n}{\eta_0(K+1)^2}T_0 + \frac{\eta_0\rho_Q\tilde{\sigma}^2(K+1)}{n}.$$

Optimizing the right hand side with respect to η_0 under the constraint $\eta_0 \leq \frac{1}{3L_Q}$ gives (34) and it is then easy to see that (35) is satisfied. ☐