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

A. Making SAGA Robust to Stochastic Perturbations

Algorithm 3 Iteration (A) with SAGA estimator

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

B. Details about the Experimental Setup

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

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

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

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

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

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

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

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

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

C. Useful Mathematical Results

C.1. Simple Results about Convexity and Smoothness

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

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

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

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

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

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

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

C.2. Useful Results to Select Step Sizes

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

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

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

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

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

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

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

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

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

Besides,

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

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

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

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

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

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

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

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

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

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

C.3. Averaging Strategy

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

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

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

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

No averaging.

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

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

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

then,

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

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

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

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

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

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

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

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

D. Proofs of the Main Results

D.1. Proof of Proposition 1

Proof.

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

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

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

where we have defined the following quantity

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

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

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

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

D.2. Proof of Corollary 2

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

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

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

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

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

D.3. Proof of Proposition 2

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

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

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

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

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

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

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

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

D.4. Proof of Proposition 3

Proof. To make the notation more compact, we call

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

Then, according to Proposition 2, we have

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

and according to Proposition 1,

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

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

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

By taking a weighted average, this yields

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

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

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

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

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

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

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

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

D.5. Proof of Corollary 3

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

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

Then, note that

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

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

D.6. Proof of Corollary 4

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

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

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

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

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

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

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

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

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

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

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

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

which gives the desired complexity.

□

D.7. Proof of Theorem 3

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

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

Then, we characterize the quantity d_k^*:

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

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

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

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

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

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

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

We may then use Lemma 2, which gives us

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

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

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

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

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

D.8. Proof of Lemma 1

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

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

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

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

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

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

Then, it is easy to show that

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

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

D.9. Proof of Lemma 2

Proof.

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

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

D.10. Proof of Corollary 6

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

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

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

D.11. Proof of Proposition 4

Proof.

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

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

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

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). □

D.12. Proof of Lemma 3

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

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

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. □

D.13. Proof of Theorem 4

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

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

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

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

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,

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

We may now use Lemma 3, which gives us

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

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

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

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

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

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

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

and we obtain the desired result. □

D.14. Proof of Corollary 8

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

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

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

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