Appendix

This supplementary material is organized as follows:

- In Appendix A we discuss some fundamental properties of zero-order gradient estimation techniques.
- In Appendix B we give detailed proofs for the results in Section 4.
- In Appendix C we give detailed proofs for the results in Section 5.
- In Appendix D we give further details for the experiments in Section 6 and provide additional empirical results.

We summarize some notation used in the paper. The vector $x \in \mathbb{R}^d$ is the variable to optimize and n is the cardinality of the dataset. The Gaussian smoothed gradient estimator is denoted as g_μ (Eq. (2)) while the coordinate-wise gradient estimator is denoted as g_ν (Eq. (3)). The variable i indexes the data-point, and sometimes we specify it in the subscript of our estimators, e.g. $g_{\mu,i}, g_{\nu,i}$. The vector $u \in \mathbb{R}^d$ is the random direction generated from $\mathcal{N}(0, I_d)$ for Gaussian smoothing estimator (I_d is the identity matrix in $\mathbb{R}^d$). D_0, D_0' are some suboptimality measures for the initial states (see main paper). s is the index of epoch, and we usually omit the superscript when discussing inner iterations inside each epoch, e.g. x_t (which should be denoted as x_t^s rigorously). The pivotal information always has a superscript "$\sim$", e.g. $\tilde{x}$ denotes the current pivotal point and $\tilde{g}$ denotes the pivotal gradient estimation at epoch s.

A. Zero-Order Gradient estimation with variance reduction

We discuss here some fundamental properties of zero-order gradient estimation. We will use these properties heavily in Appendix B and Appendix C.

A.1. Gaussian smoothing approach

We start by recalling some definitions presented in Section 3 of the main paper. Consider a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$; its smoothed version $f_\mu : \mathbb{R}^d \to \mathbb{R}$ is defined pointwise as

$$f_\mu(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(x + \mu u) e^{-\frac{1}{2}\|u\|^2} du, \quad \forall x \in \mathbb{R}^d.$$

We list some useful properties of f_μ in the next lemma.
We recall that we say f is *L-smooth* if, $\forall x, y \in \mathbb{R}^d$, $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$.

Lemma 10. *The following properties hold :*
(1) If f is convex, then f_μ is also convex.

(2) If f is L-smooth, then f_μ is also L-smooth.

(3) If f is τ-strongly convex, then f_μ is also τ-strongly convex.

(4) (Lemma 1 in (Nesterov & Spokoiny, 2011)) Let $u \sim \mathcal{N}(0, I_d)$, the standard normal distribution in $\mathbb{R}^d$. For $p \ge 2$, $d^{\frac{p}{2}} \le \mathbb{E}_u\big[\|u\|^p\big] \le (d+p)^{\frac{p}{2}}$.

We give a proof of the third property, since it is not explicitly carried out in (Nesterov & Spokoiny, 2011).

Proof of Lemma 10. f is τ-strongly convex if and only if (see e.g. Theorem 2.1.9 in (Nesterov, 2014)) for all $x', y' \in \mathbb{R}^d$ and $\alpha \in [0, 1]$,

$$f(\alpha x' + (1-\alpha)y') \le \alpha f(x') + (1-\alpha) f(y') - \frac{\alpha(1-\alpha)\tau}{2}\|x' - y'\|^2.$$

We want to prove the same inequality for f_μ. Let $x, y \in \mathbb{R}^d$ and $\alpha \in [0,1]$:

$$\begin{aligned} f_\mu(\alpha x + (1-\alpha)y) &= \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(\alpha x + (1-\alpha)y + \mu u) e^{-\frac{1}{2}\|u\|^2} du \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(\alpha(x+\mu u) + (1-\alpha)(y+\mu u)) e^{-\frac{1}{2}\|u\|^2} du. \end{aligned}$$

By picking $x' = x + \mu u$ and $y' = y + \mu u$ in the definition of strong convexity for f, by linearity of integration and noting that $x' - y' = x - y$, we get the desired result:

$$\begin{aligned} f_\mu(\alpha x + (1-\alpha)y) &\le \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \left(\alpha f(x+\mu u) + (1-\alpha) f(y+\mu u) - \frac{\alpha(1-\alpha)\tau}{2}\|x-y\|^2 \right) e^{-\frac{1}{2}\|u\|^2} du \\ &= \alpha f_\mu(x) + (1-\alpha) f_\mu(y) - \frac{\alpha(1-\alpha)\tau}{2}\|x-y\|^2, \end{aligned}$$

where in the last equality we used the fact that $\frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} e^{-\frac{1}{2}\|u\|^2} du = 1$. □

Properties of the smoothed gradient field. Note that $f_\mu(x) = \mathbb{E}_u[f(x+\mu u)]$, with $u \sim \mathcal{N}(0, I_d)$, the standard normal distribution in $\mathbb{R}^d$. Hence, the gradient of f_μ can be written as

$$\nabla f_\mu(x) = \mathbb{E}_u[g_\mu(x,u)], \quad g_\mu(x,u) := \frac{(f(x+\mu u) - f(x))u}{\mu}.$$

We list below some useful bounds from (Nesterov & Spokoiny, 2011).

Lemma 11. *If $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth, then*

(1) (Theorem 1 from (Nesterov & Spokoiny, 2011))

$$|f_\mu(x) - f(x)| \le \frac{\mu^2 L d}{2};$$

and, if f is convex (inequality (11) in (Nesterov & Spokoiny, 2011))

$$f_\mu(x) \ge f(x);$$

(2) (Lemma 3 from (Nesterov & Spokoiny, 2011))

$$\|\nabla f_\mu(x) - \nabla f(x)\| \le \frac{\mu L (d+3)^{\frac{3}{2}}}{2};$$

(3) (Lemma 4 from (Nesterov & Spokoiny, 2011))

$$\|\nabla f(x)\|^2 \le 2\|\nabla f_\mu(x)\|^2 + \frac{\mu^2 L^2 (d+6)^3}{2};$$

(4) (Theorem 4 from (Nesterov & Spokoiny, 2011))

$$\mathbb{E}_u\left[\|g_\mu(x,u)\|^2\right] \le \frac{\mu^2 L^2 (d+6)^3}{2} + 2(d+4)\|\nabla f(x)\|^2;$$

(5) (Lemma 5 from (Nesterov & Spokoiny, 2011))

$$\mathbb{E}_u\left[\|g_\mu(x,u)\|^2\right] \le 3\mu^2 L^2 (d+4)^3 + 4(d+4)\|\nabla f_\mu(x)\|^2.$$

Stochastic approximation of ∇f_μ. In the context of this paper, $f := \frac{1}{n}\sum_i f_i$. A stochastic estimate of $g_\mu(x,u)$ using data-point i can be then calculated as follows:

$$g_\mu(x,u,i) := \frac{f_i(x+\mu u) - f_i(x)}{\mu} u, \quad u \sim \mathcal{N}(0, I_d).$$

In the inner loop of Algorithm 1, at iteration t, we use $g_\mu(x,u,i)$ to get a *variance-reduced gradient estimate* of $\nabla f_\mu(\underline{x}_t)$:

$$G_t := g_\mu(\underline{x}_t, u_t, i_t) - g_\mu(\tilde{x}, u_t, i_t) + \tilde{g},$$

where *we dropped the epoch index* (i.e. s) for simplicity, as we will often do in the next pages. To study Algorithm 1, it is necessary to get an estimate of $\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\|G_t - \mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}[G_t]\|^2\right]$, where $\mathcal{F}_{t-1}$ denotes the past iterates in the current epoch. Such a bound is provided by Lemma 1 — our main lemma for DFO variance reduction. Before proving this bound, we need a result from (Nesterov & Spokoiny, 2011).

Lemma 12. *(Theorem 3 from (Nesterov & Spokoiny, 2011)) Denote $f'(x,u)$ the directional derivative of f at x along direction u:*

$$f'(x,u) = \lim_{\alpha\downarrow 0} \frac{1}{\alpha}\big[f(x+\alpha u) - f(x)\big].$$

Let $g_0(x,u) := f'(x,u)\cdot u$. If f is differentiable at x, then $f'(x,u) = \langle \nabla f(x), u\rangle$, and $g_0(x,u) = \langle \nabla f(x), u\rangle \cdot u$. Also, the following inequality holds:

$$\mathbb{E}_u\left[\|g_0(x,u)\|^2\right] \le (d+4)\|\nabla f(x)\|^2.$$

Now, let us start the proof of Lemma 1 in the main paper. *Note that this lemma requires each f_i to be L-smooth.*

Proof of Lemma 1. According to $\mathbb{E}\left[||\xi - \mathbb{E}[\xi]||^2\right] \le \mathbb{E}\left[||\xi||^2\right]$, we have

$$\begin{aligned}\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\|G_t - \mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}[G_t]\|^2\right] &= \mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\|g_\mu(\underline{x}_t,u_t,i_t) - g_\mu(\tilde{x},u_t,i_t) - \nabla f_\mu(\underline{x}_t) + \nabla f_\mu(\tilde{x})\|^2\right]\\ &\le \mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\|g_\mu\left(\underline{x}_t,u_t,i_t\right) - g_\mu\left(\tilde{x},u_t,i_t\right)\|^2\right].\end{aligned}$$

The term $\|g_\mu\left(\underline{x}_t,u_t,i_t\right) - g_\mu\left(\tilde{x},u_t,i_t\right)\|^2$ can be bounded as follows:

$$\begin{aligned}&\|g_\mu\left(\underline{x}_t,u_t,i_t\right) - g_\mu\left(\tilde{x},u_t,i_t\right)\|^2\\ =&\left\|\frac{f_{i_t}(\underline{x}_t+\mu u_t) - f_{i_t}(\underline{x}_t)}{\mu}\cdot u_t - \frac{f_i(\tilde{x}+\mu u_t) - f_{i_t}(\tilde{x})}{\mu}\cdot u_t\right\|^2\\ =&\left\|\frac{f_{\mu,i_t}(\underline{x}_t+\mu u_t) - e_1 - f_{\mu,i_t}(\underline{x}_t) + e_2}{\mu}\cdot u_t - \frac{f_{\mu,i}(\tilde{x}+\mu u_t) - e_3 - f_{\mu,i_t}(\tilde{x}) + e_4}{\mu}\cdot u_t\right\|^2,\end{aligned}$$

where e_1, e_2, e_3, e_4 denote some errors due to the *small* difference between f_{i_t} and f_{μ,i_t}, which we will bound shortly. We proceed with some additional algebraic manipulations.

$$\begin{aligned}&\|g_\mu\left(\underline{x}_t,u_t,i_t\right) - g_\mu\left(\tilde{x},u_t,i_t\right)\|^2\\ =&\left\|\frac{f_{\mu,i_t}(\underline{x}_t+\mu u_t) - f_{\mu,i_t}(\underline{x}_t)}{\mu}\cdot u_t - \frac{f_{\mu,i}(\tilde{x}+\mu u_t) - f_{\mu,i_t}(\tilde{x})}{\mu}\cdot u_t + \frac{e_1-e_2-e_3+e_4}{\mu}\cdot u_t\right\|^2\\ =&\left\|\frac{f_{\mu,i_t}(\underline{x}_t+\mu u_t) - f_{\mu,i_t}(\underline{x}_t) - \mu\langle\nabla f_{\mu,i_t}(\underline{x}_t), u_t\rangle}{\mu}\cdot u_t - \frac{f_{\mu,i_t}(\tilde{x}+\mu u_t) - f_{\mu,i_t}(\tilde{x}) - \mu\langle\nabla f_{\mu,i_t}(\tilde{x}), u_t\rangle}{\mu}\cdot u_t\right.\\ &\left.+\langle\nabla f_{\mu,i_t}(\underline{x}_t) - \nabla f_{\mu,i_t}(\tilde{x}), u_t\rangle\cdot u_t + + \frac{e_1-e_2-e_3+e_4}{\mu}\cdot u_t\right\|^2\\ \le& 4\Bigg(\left\|\frac{f_{\mu,i_t}(\underline{x}_t+\mu u_t) - f_{\mu,i_t}(\underline{x}_t) - \mu\langle\nabla f_{\mu,i_t}(\underline{x}_t), u_t\rangle}{\mu}\cdot u_t\right\|^2 + \left\|\frac{f_{\mu,i_t}(\tilde{x}+\mu u_t) - f_{\mu,i_t}(\tilde{x}) - \mu\langle\nabla f_{\mu,i_t}(\tilde{x}), u_t\rangle}{\mu}\cdot u_t\right\|^2\\ &+\|\langle\nabla f_{\mu,i_t}(\underline{x}_t) - \nabla f_{\mu,i_t}(\tilde{x}), u_t\rangle\cdot u_t\|^2 + \left\|\frac{e_1-e_2-e_3+e_4}{\mu}\cdot u_t\right\|^2\Bigg)\\ \le& 4\Bigg(\left(\frac{\mu}{2}L\|u_t\|^3\right)^2 + \left(\frac{\mu}{2}L\|u_t\|^3\right)^2 + \|\langle\nabla f_{\mu,i_t}(\underline{x}_t) - \nabla f_{\mu,i_t}(\tilde{x}), u_t\rangle\cdot u_t\|^2 + \left\|\frac{e_1-e_2-e_3+e_4}{\mu}\cdot u_t\right\|^2\Bigg)\end{aligned}$$

$$\leq 4\left(\frac{\mu^2}{2}L^2\|u_t\|^6 + \|\langle \nabla f_{\mu,i_t}(\underline{x}_t) - \nabla f_{\mu,i_t}(\tilde{x}), u_t\rangle \cdot u_t\|^2 + 4\mu^2 L^2 d^2 \|u_t\|^2\right),$$

where the second last inequality comes from the smoothness of $f_{\mu,i}$ and the last inequality is from (1) in Lemma 11. Now, we define a new function $f^e_{\mu,i_t}(x) = f_{\mu,i_t}(x) - \langle \nabla f_{\mu,i_t}(\tilde{x}), x\rangle$ and

$$\nabla f^e_{\mu,i_t}(x) = \nabla f_{\mu,i_t}(x) - \nabla f_{\mu,i_t}(\tilde{x}).$$

Also, we define $g^e_{0,\mu,i_t}(x,u)$ as

$$g^e_{0,\mu,i_t}(x,u) := \left(f^e_{\mu,i_t}\right)'(x,u) \cdot u. \tag{19}$$

Note that $g^e_{0,\mu,i_t}(x,u_t)$ is related to the second term in the inequality before:

$$\|g^e_{0,\mu,i_t}(x,u_t)\|^2 = \langle \nabla f^e_{\mu,i_t}(x), u_t\rangle^2 \cdot \|u_t\|^2 = \|\langle \nabla f_{\mu,i_t}(x) - \nabla f_{\mu,i_t}(\tilde{x}), u_t\rangle \cdot u_t\|^2.$$

Next, we apply Lemma 12 to f^e_{μ,i_t}:

$$\begin{aligned}\mathbb{E}_{u_t}\left[\|g^e_{0,\mu,i_t}(x,u_t)\|^2\right] &\leq (d+4)\|\nabla f^e_{\mu,i_t}(x)\|^2 \\ &= (d+4)\|\nabla f_{\mu,i_t}(x) - \nabla f_{\mu,i_t}(\tilde{x})\|^2.\end{aligned}$$

Putting it all together, we obtain the desired bound:

$$\begin{aligned}&\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\|G_t - \mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}[G_t]\|^2\right] \\ &\leq \mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\|g_\mu\left(\underline{x}_t, u_t, i_t\right) - g_\mu\left(\tilde{x}, u_t, i_t\right)\|^2\right] \\ &\leq 2\mu^2 L^2 \mathbb{E}_{u_t|\mathcal{F}_{t-1}}\left[\|u_t\|^6\right] + 4(d+4)\mathbb{E}_{i_t|\mathcal{F}_{t-1}}\left[\|\nabla f_{\mu,i_t}(\underline{x}_t) - \nabla f_{\mu,i_t}(\tilde{x})\|^2\right] + 16\mu^2 L^2 d^2 \mathbb{E}_{u_t|\mathcal{F}_{t-1}}\left[\|u_t\|^2\right] \\ &\leq 2\mu^2 L^2 (d+6)^3 + 16\mu^2 L^2 d^3 + 8(d+4)L\mathbb{E}_{i_t|\mathcal{F}_{t-1}}\left[f_{\mu,i_t}(\tilde{x}) - f_{\mu,i_t}(\underline{x}_t) - \langle \nabla f_{\mu,i_t}(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\right] \\ &\leq 18\mu^2 L^2 (d+6)^3 + 8(d+4)L\left[f_\mu(\tilde{x}) - f_\mu(\underline{x}_t) - \langle \nabla f_\mu(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\right].\end{aligned}$$

The second last inequality holds thanks to Theorem 2.1.5 in (Nesterov, 2014). □

A.2. Coordinate-wise approach

In Section 5 we replace the Gaussian smoothing estimator of Eq. (2) with the coordinate-wise approach of (Ji et al., 2019) for computing G_t in Algorithm 1. That is, we set

$$G_t = g_\nu(\underline{x}_t, i_t) - g_\nu(\tilde{x}, i_t) + g_\nu(\tilde{x}),$$

with, as we specified in Eq. (3) of the main paper:

$$g_\nu(x,i) = \sum_{j=1}^{d} \frac{f_i(x + \nu \mathrm{e}_j) - f_i(x - \nu \mathrm{e}_j)}{2\nu}\mathrm{e}_j,$$

where e_j is the unit vector with only one non-zero entry 1 at its j^{th} coordinate. Note that, *g_ν is d times more expensive to compute compared compared to g_μ*, which we discussed before.
The following lemma gives an useful approximation error bound.

Lemma 13. *(Lemma 3 (Appendix D) from (Ji et al., 2019)) Suppose each f_i is L-smooth and that we use the coordinate-wise gradient estimation in Eq. (3). For any smoothing parameter $\nu > 0$ and any $x \in \mathbb{R}^d$, we have*

$$\|g_\nu(x,i) - \nabla f_i(x)\|^2 \leq L^2 d\nu^2.$$

Also, if we define $g_\nu(x) := \frac{1}{n}\sum_{i=1}^{n} g_\nu(x,i)$, we clearly have $\|g_\nu(x) - \nabla f(x)\|^2 \leq L^2 d\nu^2$.

In the next lemma, we bound the variance of G_t. As the reader will soon notice, compared to Lemma 1, the proof in the coordinate-wise case is simpler and closely related to the standard variance reduction analysis [5].

[5] See e.g. Lemma 2.4 in (Allen-Zhu, 2017).

Lemma 14. *When we use coordinate-wise gradient estimator Eq. (3) for computing G_t, we can obtain a DFO variance reduction as follows:*

$$\mathbb{E}_{i_t}\left[\|G_t - \nabla f(\underline{x}_t)\|^2\right] \leq 12L^2 d\nu^2 + 8L\big[f(\tilde{x}) - f(\underline{x}_t) - \langle \nabla f(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big], \tag{20}$$

where G_t is defined as

$$G_t = g_\nu(\underline{x}_t, i_t) - g_\nu(\tilde{x}, i_t) + g_\nu(\tilde{x})$$

and g_ν is the gradient estimator as defined by Eq. (3). Moreover, the expectation of the gradient estimation is

$$\mathbb{E}_{i_t}\big[\delta_t\big] = g_\nu(\underline{x}_t) - \nabla f(\underline{x}_t) \neq 0, \tag{21}$$

which is different from $\mathbb{E}_{i_t|\mathcal{F}_{t-1}}\big[\delta_t\big] = g_\nu(\tilde{x}) - \nabla f_\mu(\tilde{x})$ in Lemma 1.

Proof. Note that $G_t - \nabla f(\underline{x}_t)$ can be decoupled as

$$\begin{aligned} G_t - \nabla f(\underline{x}_t) = \nabla f_{i_t}(\underline{x}_t) - \nabla f_{i_t}(\tilde{x}) - \big(\nabla f(\underline{x}_t) - \nabla f(\tilde{x})\big) + g_\nu(\underline{x}_t, i_t) - \nabla f_{i_t}(\underline{x}_t) \\ - g_\nu(\tilde{x}, i_t) + \nabla f_{i_t}(\tilde{x}) + g_\nu(\tilde{x}) - \nabla f(\tilde{x}). \end{aligned}$$

Therefore, we have

$$\begin{aligned} \mathbb{E}_{i_t}\left[\|G_t - \nabla f(\underline{x}_t)\|^2\right] &\leq 4\mathbb{E}_{i_t}\Big[\|\nabla f_{i_t}(\underline{x}_t) - \nabla f_{i_t}(\tilde{x}) - \big(\nabla f(\underline{x}_t) - \nabla f(\tilde{x})\big)\|^2 + \|g_\nu(\underline{x}_t, i_t) - \nabla f_{i_t}(\underline{x}_t)\|^2 \\ &\qquad + \|g_\nu(\tilde{x}, i_t) - \nabla f_{i_t}(\tilde{x})\|^2 + \|g_\nu(\tilde{x}) - \nabla f(\tilde{x})\|^2\Big] \\ &\leq 4\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(\underline{x}_t) - \nabla f_{i_t}(\tilde{x})\|^2 + 3L^2 d\nu^2\right] \\ &\leq 8L\mathbb{E}_{i_t}\left[f_{i_t}(\tilde{x}) - f_{i_t}(\underline{x}_t) - \langle \nabla f_{i_t}(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\right] + 12L^2 d\nu^2 \\ &= 8L\big[f(\tilde{x}) - f(\underline{x}_t) - \langle \nabla f(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] + 12L^2 d\nu^2. \end{aligned}$$

The second inequality holds because of $\mathbb{E}\big[\|\xi - \mathbb{E}[\xi]\|^2\big] \leq \mathbb{E}\big[\|\xi\|^2\big]$ and thanks to Lemma 13. The last inequality holds thanks to Theorem 2.1.5 in (Nesterov, 2014). □

B. Proofs for Section 4

The proofs of Theorem 2 and Theorem 4 follow the same structure as in (Lan et al., 2019), with some modifications due to the zero-order gradient estimation techniques, supported by the bounds in Appendix A. We recall our basic assumption:

(A1) Each f_i is convex, differentiable and L-smooth. Hence, also $f = \frac{1}{n}\sum_{i=1}^n f_i$ is convex and L-smooth.

To make the notation compact, we define, again in analogy with (Lan et al., 2019):

$$x_{t-1}^+ := \frac{1}{1+\tau\gamma_s}(x_{t-1} + \tau\gamma_s \underline{x}_t) \tag{22}$$

and

$$l_f(z, x) := f(z) + \langle \nabla f(z), x - z\rangle. \tag{23}$$

Using the definition of $\bar{x}_t$ and x_t in Algorithm 1, we have:

$$\bar{x}_t - \underline{x}_t = \alpha_s(x_t - x_{t-1}^+). \tag{24}$$

The first result is simply an adaptation of Lemma 5 in (Lan et al., 2019) for the non-regularized Euclidean case. Hence, it does not require a proof.

Lemma 15. *Assume **(A1)**. For any $x \in \mathbb{R}^d$, we have*

$$\gamma_s[l_{f_\mu}(\underline{x}_t, x_t) - l_{f_\mu}(\underline{x}_t, x)] \leq \\ \frac{\tau\gamma_s}{2}\|\underline{x_t} - x\|^2 + \frac{1}{2}\|x_{t-1} - x\|^2 - \frac{1+\tau\gamma_s}{2}\|x_t - x\|^2 - \frac{1+\tau\gamma_s}{2}\|x_t - x_{t-1}^+\|^2 - \gamma_s\langle\delta_t, x_t - x\rangle,$$

which can be rewritten as

$$\gamma_s\langle\nabla f_\mu(\underline{x}_t), x_t - x\rangle \leq \frac{\tau\gamma_s}{2}\|\underline{x_t} - x\|^2 + \frac{1}{2}\|x_{t-1} - x\|^2 - \frac{1+\tau\gamma_s}{2}\|x_t - x\|^2 - \frac{1+\tau\gamma_s}{2}\|x_t - x_{t-1}^+\|^2 - \gamma_s\langle\delta_t, x_t - x\rangle.$$

The following lemma bounds the progress made at each inner iteration, and is similar to Lemma 6 in (Lan et al., 2019), but with some additional error terms coming from the zero-order estimation error for the gradients.

Lemma 16. *Assume **(A1)**. Assume that $\alpha_s \in [0, 1]$, $p_s \in [0, 1]$ and $\gamma_s > 0$ satisfy*

$$1 + \tau\gamma_s - L\alpha_s\gamma_s > 0, \tag{25}$$

$$p_s - \frac{4(d+4)L\alpha_s\gamma_s}{1 + \tau\gamma_s - L\alpha_s\gamma_s} \geq 0. \tag{26}$$

Conditioned on past events $\mathcal{F}_{t-1}$ and taking the expectation of u_t, i_t, we have

$$\begin{aligned}
&\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\frac{\gamma_s}{\alpha_s}\left[f_\mu(\bar{x}_t) - f_\mu(x)\right] + \frac{(1+\tau\gamma_s)}{2}\|x_t - x\|^2\right] \\
&\leq \frac{\gamma_s}{\alpha_s}(1 - \alpha_s - p_s)\left[f_\mu(\bar{x}_{t-1}) - f_\mu(x)\right] + \frac{\gamma_s p_s}{\alpha_s}\left[f_\mu(\tilde{x}) - f_\mu(x)\right] + \frac{1}{2}\|x_{t-1} - x\|^2 + \frac{\gamma_s}{\alpha_s}\cdot\frac{9\alpha_s\gamma_s\mu^2L^2(d+6)^3}{1 + \tau\gamma_s - L\alpha_s\gamma_s} \\
&\quad - \frac{\gamma_s}{\alpha_s}\cdot\alpha_s\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\langle\tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x\rangle\right]
\end{aligned} \tag{27}$$

for any $x \in \mathbb{R}^d$.

Remark. *The second term in Eq. (26) has a dependency on $(d + 4)$, due to the Gaussian smoothing distortion.*

Proof of Lemma 16. By the L-smoothness of f_μ (from Lemma 10),

$$\begin{aligned}
f_\mu(\bar{x}_t) &\leq f_\mu(\underline{x}_t) + \langle\nabla f_\mu(\underline{x}_t), \bar{x}_t - \underline{x}_t\rangle + \frac{L}{2}\|\bar{x}_t - \underline{x}_t\|^2 \\
&= (1 - \alpha_s - p_s)\left[f_\mu(\underline{x}_t) + \langle\nabla f_\mu(\underline{x}_t), \bar{x}_{t-1} - \underline{x}_t\rangle\right] + \alpha_s\left[f_\mu(\underline{x}_t) + \langle\nabla f_\mu(\underline{x}_t), x_t - \underline{x}_t\rangle\right] \\
&\quad + p_s\left[f_\mu(\underline{x}_t) + \langle\nabla f_\mu(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\right] + \frac{L\alpha_s^2}{2}\|x_t - x_{t-1}^+\|^2.
\end{aligned}$$

The equality above holds because of the update rule of $\bar{x}_t$ in Algorithm 1 and the Eq. (24). Next, applying Lemma 15 for the inequality above, we have

$$\begin{aligned}
f_\mu(\bar{x}_t) \leq{}& (1 - \alpha_s - p_s)\left[f_\mu(\underline{x}_t) + \langle\nabla f_\mu(\underline{x}_t), \bar{x}_{t-1} - \underline{x}_t\rangle\right] + \alpha_s\left[f_\mu(\underline{x}_t) + \langle\nabla f_\mu(\underline{x}_t), x - \underline{x}_t\rangle\right] \\
&+ \alpha_s\left[\frac{\tau}{2}\|\underline{x_t} - x\|^2 + \frac{1}{2\gamma_s}\|x_{t-1} - x\|^2 - \frac{1+\tau\gamma_s}{2\gamma_s}\|x_t - x\|^2 - \frac{1+\tau\gamma_s}{2\gamma_s}\|x_t - x_{t-1}^+\|^2 - \langle\delta_t, x_t - x\rangle\right] \\
&+ p_s\left[f_\mu(\underline{x}_t) + \langle\nabla f_\mu(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\right] + \frac{L\alpha_s^2}{2}\|x_t - x_{t-1}^+\|^2 \\
\leq{}& (1 - \alpha_s - p_s)\left[f_\mu(\bar{x}_{t-1}) - \frac{\tau}{2}\|\bar{x}_{t-1} - \underline{x}_t\|^2\right] + \alpha_s\left[f_\mu(x) - \frac{\tau}{2}\|x - \underline{x}_t\|^2\right] \\
&+ \alpha_s\left[\frac{\tau}{2}\|\underline{x_t} - x\|^2 + \frac{1}{2\gamma_s}\|x_{t-1} - x\|^2 - \frac{1+\tau\gamma_s}{2\gamma_s}\|x_t - x\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
&+ p_s\big[f_\mu(\underline{x}_t) + \langle \nabla f_\mu(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] - \frac{\alpha_s}{2\gamma_s}(1+\tau\gamma_s - L\alpha_s\gamma_s)\|x_t - x_{t-1}^{+}\|^2 - \alpha_s\langle \delta_t, x_t - x\rangle \\
&= (1-\alpha_s-p_s)\big[f_\mu(\bar{x}_{t-1}) - \frac{\tau}{2}\|\bar{x}_{t-1} - \underline{x}_t\|^2\big] + \alpha_s\big[f_\mu(x) + \frac{1}{2\gamma_s}\|x_{t-1} - x\|^2 - \frac{1+\tau\gamma_s}{2\gamma_s}\|x_t - x\|^2\big] \\
&\quad + p_s\big[f_\mu(\underline{x}_t) + \langle \nabla f_\mu(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] - \frac{\alpha_s}{2\gamma_s}(1+\tau\gamma_s - L\alpha_s\gamma_s)\|x_t - x_{t-1}^{+}\|^2 \\
&\quad - \alpha_s\langle \delta_t - \tilde{g} + \nabla f_\mu(\tilde{x}), x_t - x_{t-1}^{+}\rangle - \alpha_s\langle \tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x_{t-1}^{+}\rangle - \alpha_s\langle \delta_t, x_{t-1}^{+} - x\rangle \\
&= (1-\alpha_s-p_s)\big[f_\mu(\bar{x}_{t-1}) - \frac{\tau}{2}\|\bar{x}_{t-1} - \underline{x}_t\|^2\big] + \alpha_s\big[f_\mu(x) + \frac{1}{2\gamma_s}\|x_{t-1} - x\|^2 - \frac{1+\tau\gamma_s}{2\gamma_s}\|x_t - x\|^2\big] \\
&\quad + p_s\big[f_\mu(\underline{x}_t) + \langle \nabla f_\mu(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] - \frac{\alpha_s}{2\gamma_s}(1+\tau\gamma_s - L\alpha_s\gamma_s)\|x_t - x_{t-1}^{+}\|^2 \\
&\quad - \alpha_s\langle g_\mu(\underline{x}_t, u_t, i_t) - g_\mu(\tilde{x}, u_t, i_t) - \nabla f_\mu(\underline{x}_t) + \nabla f_\mu(\tilde{x}), x_t - x_{t-1}^{+}\rangle \\
&\quad - \alpha_s\langle \tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x_{t-1}^{+}\rangle - \alpha_s\langle \delta_t, x_{t-1}^{+} - x\rangle.
\end{aligned}
$$

The second inequality holds thanks to the strong convexity (with $\tau \geq 0$) of f_μ (see again Lemma 10) and the last equality comes from the definition

$$\delta_t = G_t - \nabla f_\mu(\underline{x}_t).$$

Next, note that for any $a > 0, b \in \mathbb{R}$ and $u, v \in \mathbb{R}^d$, it holds that $b\langle u, v\rangle - \frac{a}{2}\|v\|^2 \leq \frac{b^2}{2a}\|u\|^2$. If we set $a = \frac{\alpha_s}{\gamma_s}(1 + \tau\gamma_s - L\alpha_s\gamma_s)$ and $b = -\alpha_s$ (requiring $1 + \tau\gamma_s - L\alpha_s\gamma_s > 0$), we get

$$
\begin{aligned}
f_\mu(\bar{x}_t) &\leq (1-\alpha_s-p_s)\big[f_\mu(\bar{x}_{t-1}) - \frac{\tau}{2}\|\bar{x}_{t-1} - \underline{x}_t\|^2\big] + \alpha_s\big[f_\mu(x) + \frac{1}{2\gamma_s}\|x_{t-1} - x\|^2 - \frac{1+\tau\gamma_s}{2\gamma_s}\|x_t - x\|^2\big] \\
&\quad + p_s\big[f_\mu(\underline{x}_t) + \langle \nabla f_\mu(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] + \frac{\alpha_s\gamma_s}{2(1+\tau\gamma_s - L\alpha_s\gamma_s)}\|g_\mu(\underline{x}_t, u_t, i_t) - g_\mu(\tilde{x}, u_t, i_t) - \nabla f_\mu(\underline{x}_t) + \nabla f_\mu(\tilde{x})\|^2 \\
&\quad - \alpha_s\langle \tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x_{t-1}^{+}\rangle - \alpha_s\langle \delta_t, x_{t-1}^{+} - x\rangle \\
&= (1-\alpha_s-p_s)\big[f_\mu(\bar{x}_{t-1}) - \frac{\tau}{2}\|\bar{x}_{t-1} - \underline{x}_t\|^2\big] + \alpha_s\big[f_\mu(x) + \frac{1}{2\gamma_s}\|x_{t-1} - x\|^2 - \frac{1+\tau\gamma_s}{2\gamma_s}\|x_t - x\|^2\big] \\
&\quad + p_s\big[f_\mu(\underline{x}_t) + \langle \nabla f_\mu(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] + \frac{\alpha_s\gamma_s}{2(1+\tau\gamma_s - L\alpha_s\gamma_s)}\|G_t - \mathbb{E}_{u_t, i_t}[G_t]\|^2 \\
&\quad - \alpha_s\langle \tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x_{t-1}^{+}\rangle - \alpha_s\langle \delta_t, x_{t-1}^{+} - x\rangle. \qquad (28)
\end{aligned}
$$

Taking the expectation w.r.t. u_t, i_t conditional on past iterates and applying Lemma 1,

$$
\begin{aligned}
&p_s\big[f_\mu(\underline{x}_t) + \langle \nabla f_\mu(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] + \frac{\alpha_s\gamma_s}{2(1+\tau\gamma_s - L\alpha_s\gamma_s)}\mathbb{E}_{u_t, i_t|\mathcal{F}_{t-1}}\big[\|G_t - \mathbb{E}_{u_t, i_t}[G_t]\|^2\big] \\
&\quad - \alpha_s\mathbb{E}_{u_t, i_t|\mathcal{F}_{t-1}}\big[\langle \tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x_{t-1}^{+}\rangle\big] - \alpha_s\mathbb{E}_{u_t, i_t|\mathcal{F}_{t-1}}\big[\langle \delta_t, x_{t-1}^{+} - x\rangle\big] \\
&\leq p_s\big[f_\mu(\underline{x}_t) + \langle \nabla f_\mu(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] + \frac{9\alpha_s\gamma_s \cdot \mu^2 L^2 (d+6)^3}{1+\tau\gamma_s - L\alpha_s\gamma_s} + \frac{4\alpha_s\gamma_s(d+4)L}{1+\tau\gamma_s - L\alpha_s\gamma_s}\big[f_\mu(\tilde{x}) - f_\mu(\underline{x}_t) - \langle \nabla f_\mu(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] \\
&\quad - \alpha_s\mathbb{E}_{u_t, i_t|\mathcal{F}_{t-1}}\big[\langle \tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x\rangle\big] \\
&= \big(p_s - \frac{4\alpha_s\gamma_s(d+4)L}{1+\tau\gamma_s - L\alpha_s\gamma_s}\big)\big[f_\mu(\underline{x}_t) + \langle \nabla f_\mu(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] + \frac{9\alpha_s\gamma_s \cdot \mu^2 L^2 (d+6)^3}{1+\tau\gamma_s - L\alpha_s\gamma_s} + \frac{4\alpha_s\gamma_s(d+4)L}{1+\tau\gamma_s - L\alpha_s\gamma_s}f_\mu(\tilde{x}) \\
&\quad - \alpha_s\mathbb{E}_{u_t, i_t|\mathcal{F}_{t-1}}\big[\langle \tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x\rangle\big] \\
&\leq \big(p_s - \frac{4\alpha_s\gamma_s(d+4)L}{1+\tau\gamma_s - L\alpha_s\gamma_s}\big)\big[f_\mu(\tilde{x}) - \frac{\tau}{2}\|\tilde{x} - \underline{x}_t\|^2\big] + \frac{9\alpha_s\gamma_s \cdot \mu^2 L^2 (d+6)^3}{1+\tau\gamma_s - L\alpha_s\gamma_s} + \frac{4\alpha_s\gamma_s(d+4)L}{1+\tau\gamma_s - L\alpha_s\gamma_s}f_\mu(\tilde{x}) \\
&\quad - \alpha_s\mathbb{E}_{u_t, i_t|\mathcal{F}_{t-1}}\big[\langle \tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x\rangle\big] \\
&= p_s f_\mu(\tilde{x}) - \big(p_s - \frac{4\alpha_s\gamma_s(d+4)L}{1+\tau\gamma_s - L\alpha_s\gamma_s}\big) \cdot \frac{\tau}{2}\|\tilde{x} - \underline{x}_t\|^2 + \frac{9\alpha_s\gamma_s \cdot \mu^2 L^2 (d+6)^3}{1+\tau\gamma_s - L\alpha_s\gamma_s} \\
&\quad - \alpha_s\mathbb{E}_{u_t, i_t|\mathcal{F}_{t-1}}\big[\langle \tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x\rangle\big]. \qquad (29)
\end{aligned}
$$

The last inequality holds when $p_s - \frac{4\alpha_s\gamma_s(d+4)L}{1+\tau\gamma_s - L\alpha_s\gamma_s} \geq 0$. Combining Eq. (28) with Eq. (29), we obtain

$$\begin{aligned}
&\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[f_\mu(\bar{x}_t) + \frac{\alpha_s(1+\tau\gamma_s)}{2\gamma_s}\|x_t - x\|^2\right] \\
&\leq (1-\alpha_s-p_s)f_\mu(\bar{x}_{t-1}) + p_s f_\mu(\tilde{x}) + \alpha_s f_\mu(x) + \frac{\alpha_s}{2\gamma_s}\|x_{t-1}-x\|^2 + \frac{9\alpha_s\gamma_s\mu^2L^2(d+6)^3}{1+\tau\gamma_s - L\alpha_s\gamma_s} \\
&\quad - \alpha_s\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\langle \tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x\rangle\right] - \frac{(1-\alpha_s-p_s)\tau}{2}\|\bar{x}_{t-1} - \underline{x}_t\|^2 - \big(p_s - \frac{4\alpha_s\gamma_s(d+4)L}{1+\tau\gamma_s - L\alpha_s\gamma_s}\big)\cdot\frac{\tau}{2}\|\tilde{x} - \underline{x}_t\|^2 \\
&\leq (1-\alpha_s-p_s)f_\mu(\bar{x}_{t-1}) + p_s f_\mu(\tilde{x}) + \alpha_s f_\mu(x) + \frac{\alpha_s}{2\gamma_s}\|x_{t-1}-x\|^2 + \frac{9\alpha_s\gamma_s\mu^2L^2(d+6)^3}{1+\tau\gamma_s - L\alpha_s\gamma_s} \\
&\quad - \alpha_s\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\langle \tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x\rangle\right].
\end{aligned}$$

Multiplying both sides by $\frac{\gamma_s}{\alpha_s}$ and then rearranging the inequality, we finish the proof of this lemma, i.e. Eq. (27). □

B.1. Proof of Theorem 2

Before giving the convergence result for convex smooth f, we provide a lemma for the *epoch-wise analysis*. This lemma needs an additional technical assumption.

(A2$_\mu$) Let $x^*_\mu \in \text{argmin}_x f_\mu(x)$ and consider the sequence of approximations $\{\tilde{x}^s\}$ returned by Algorithm 1. There exist a *finite* constant $Z < \infty$, potentially dependent on L and d, such that, for μ small enough,

$$\sup_{s\geq 0} \mathbb{E}\left[\|\tilde{x}^s - x^*_\mu\|\right] \leq Z.$$

Using an argument similar to (Gadat et al., 2018), it is possible to show that this assumption holds under the requirement that f is coercive, i.e. $f(x) \to \infty$ as $\|x\| \to \infty$.

Lemma 17. *Assume (A1), (A2$_\mu$). Suppose that the weights $\{\theta_t\}$ are set as*

$$\theta_t = \begin{cases} \frac{\gamma_s}{\alpha_s}(\alpha_s + p_s) & 1 \leq t \leq T_s - 1 \\ \frac{\gamma_s}{\alpha_s} & t = T_s. \end{cases} \tag{30}$$

Define:

$$\mathcal{L}_s := \frac{\gamma_s}{\alpha_s} + (T_s - 1)\frac{\gamma_s(\alpha_s + p_s)}{\alpha_s}, \tag{31}$$

$$\mathcal{R}_s := \frac{\gamma_s}{\alpha_s}(1-\alpha_s) + (T_s - 1)\frac{\gamma_s p_s}{\alpha_s}. \tag{32}$$

Under the conditions in Eq. (25) and Eq. (26), we have:

$$\begin{aligned}
&\mathcal{L}_s\mathbb{E}_{\mathcal{F}_{T_s}}\left[f_\mu(\tilde{x}^s) - f_\mu(x^*_\mu)\right] \\
&\leq \mathcal{R}_s\cdot\left[f_\mu(\tilde{x}^{s-1}) - f_\mu(x^*_\mu)\right] + \big(\frac{1}{2}\|x^{s-1} - x^*_\mu\|^2 - \frac{1}{2}\|x^s - x^*_\mu\|^2\big) + T_s\frac{\gamma_s}{\alpha_s}\cdot\frac{9\alpha_s\gamma_s\mu^2L^2(d+6)^3}{1 - L\alpha_s\gamma_s} + (\mathcal{L}_s + \mathcal{R}_s)Z||e^s|| \\
&\leq \mathcal{R}_s\cdot\left[f_\mu(\tilde{x}^{s-1}) - f_\mu(x^*_\mu)\right] + \big(\frac{1}{2}\|x^{s-1} - x^*_\mu\|^2 - \frac{1}{2}\|x^s - x^*_\mu\|^2\big) + \underbrace{T_s\frac{\gamma_s}{\alpha_s}\cdot\frac{9\alpha_s\gamma_s\mu^2L^2(d+6)^3}{1 - L\alpha_s\gamma_s}}_{①} + \underbrace{(\mathcal{L}_s + \mathcal{R}_s)Z\sqrt{E}}_{②}.
\end{aligned}$$

where $e^s = \tilde{g}^s - \nabla f_\mu(\tilde{x}^{s-1})$ and $||e^s||^2 \leq E$, which is consistent with the definition of E in Section 4.1. Here, the expectation is taken over $\mathcal{F}_{T_s}$ inside the epoch s.

Remark. *Compared to the corresponding result by Lan et al. (2019) (Lemma 7 in their paper), we note that two additional* *errors terms appear.* $\textcircled{1}$ *is the error due to the Gaussian smooth estimation and* $\textcircled{2}$ *is an error due to the approximation made at the pivot point. We will later see that the coordinate-wise approach introduced in Eq.* (3) *yields a constant error bound for* $\textcircled{2}$*, which is independent of the gradient information.*

Proof of Lemma 17. For f convex and L-smooth, we have that f_μ is L-smooth and τ-strongly-convex with $\tau = 0$ from Lemma 10. Hence, Lemma 16 can be written as

$$\begin{aligned}
&\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\frac{\gamma_s}{\alpha_s}\big[f_\mu(\bar{x}_t) - f_\mu(x)\big] + \frac{1}{2}\|x_t - x\|^2\right] \\
&\le \frac{\gamma_s}{\alpha_s}(1-\alpha_s-p_s)\big[f_\mu(\bar{x}_{t-1}) - f_\mu(x)\big] + \frac{\gamma_s p_s}{\alpha_s}\big[f_\mu(\tilde{x}) - f_\mu(x)\big] + \frac{1}{2}\|x_{t-1} - x\|^2 + \frac{\gamma_s}{\alpha_s}\cdot\frac{9\alpha_s\gamma_s\mu^2L^2(d+6)^3}{1-L\alpha_s\gamma_s} \\
&\quad - \frac{\gamma_s}{\alpha_s}\cdot\alpha_s\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\big[\langle\tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x\rangle\big].
\end{aligned}$$

Summing up these inequalities over $t = 1, \dots, T_s$, using the definition of θ_t and $\bar{x}_0 = \tilde{x}$,

$$\begin{aligned}
\sum_{t=1}^{T_s}\theta_t\mathbb{E}_{\mathcal{F}_t}\big[f_\mu(\bar{x}_t) - f_\mu(x)\big] &\le \big[\frac{\gamma_s}{\alpha_s}(1-\alpha_s) + (T_s-1)\frac{\gamma_s p_s}{\alpha_s}\big]\cdot\big[f_\mu(\tilde{x}) - f_\mu(x)\big] + \big(\frac{1}{2}\|x_0 - x\|^2 - \frac{1}{2}\|x_{T_s} - x\|^2\big) \\
&\quad + T_s\cdot\frac{\gamma_s}{\alpha_s}\cdot\frac{9\alpha_s\gamma_s\mu^2L^2(d+6)^3}{1-L\alpha_s\gamma_s} - \frac{\gamma_s}{\alpha_s}\sum_{t=1}^{T_s}\alpha_s\mathbb{E}_{\mathcal{F}_t}\big[\langle\tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x\rangle\big].
\end{aligned}$$

Using the fact that $\tilde{x}^s = \sum_{t=1}^{T_s}\big(\theta_t\bar{x}_t\big)/\sum_{t=1}^{T_s}\theta_t$, $\tilde{x} = \tilde{x}^{s-1}$, $x_0 = x^{s-1}$, $x_{T_s} = x^s$ and using convexity of f_μ, the inequality above implies

$$\begin{aligned}
\sum_{t=1}^{T_s}\theta_t\mathbb{E}_{\mathcal{F}_t}\big[f_\mu(\tilde{x}^s) - f_\mu(x)\big] &\le \big[\frac{\gamma_s}{\alpha_s}(1-\alpha_s) + (T_s-1)\frac{\gamma_s p_s}{\alpha_s}\big]\cdot\big[f_\mu(\tilde{x}^{s-1}) - f_\mu(x)\big] + \big(\frac{1}{2}\|x^{s-1} - x\|^2 - \frac{1}{2}\|x^s - x\|^2\big) \\
&\quad + T_s\cdot\frac{\gamma_s}{\alpha_s}\cdot\frac{9\alpha_s\gamma_s\mu^2L^2(d+6)^3}{1-L\alpha_s\gamma_s} - \frac{\gamma_s}{\alpha_s}\sum_{t=1}^{T_s}\alpha_s\mathbb{E}_{\mathcal{F}_t}\big[\langle\tilde{g}^s - \nabla f_\mu(\tilde{x}^{s-1}), x_t - x\rangle\big],
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
\mathcal{L}_s\mathbb{E}_{\mathcal{F}_t}\big[f_\mu(\tilde{x}^s) - f_\mu(x)\big] &\le \mathcal{R}_s\cdot\big[f_\mu(\tilde{x}^{s-1}) - f_\mu(x)\big] + \big(\frac{1}{2}\|x^{s-1} - x\|^2 - \frac{1}{2}\|x^s - x\|^2\big) \\
&\quad + T_s\cdot\frac{\gamma_s}{\alpha_s}\cdot\frac{9\alpha_s\gamma_s\mu^2L^2(d+6)^3}{1-L\alpha_s\gamma_s} - \frac{\gamma_s}{\alpha_s}\sum_{t=1}^{T_s}\alpha_s\mathbb{E}_{\mathcal{F}_t}\big[\langle\tilde{g}^s - \nabla f_\mu(\tilde{x}^{s-1}), x_t - x\rangle\big]. \qquad (33)
\end{aligned}$$

Notice that, since $\bar{x}_0 = \tilde{x} = \tilde{x}^{s-1}$ in the epoch s,

$$\begin{aligned}
&\frac{\gamma_s}{\alpha_s}\sum_{t=1}^{T_s}\alpha_s\mathbb{E}_{\mathcal{F}_t}\big[\langle\tilde{g}^s - \nabla f_\mu(\tilde{x}^{s-1}), x_t - x\rangle\big] \\
&= \frac{\gamma_s}{\alpha_s}\sum_{t=1}^{T_s}\mathbb{E}_{\mathcal{F}_t}\big[\langle\tilde{g}^s - \nabla f_\mu(\tilde{x}^{s-1}), \bar{x}_t - (1-\alpha_s-p_s)\bar{x}_{t-1} - p_s\tilde{x}^{s-1} - \alpha_s x\rangle\big] \\
&= \frac{\gamma_s}{\alpha_s}\mathbb{E}_{\mathcal{F}_{T_s}}\big[\langle\tilde{g}^s - \nabla f_\mu(\tilde{x}^{s-1}), \bar{x}_{T_s} + \sum_{t=1}^{T_s-1}(\alpha_s+p_s)\bar{x}_t - \big[(1-\alpha_s) + (T_s-1)p_s\big]\tilde{x}^{s-1} - \alpha_s T_s x\rangle\big] \\
&= \frac{\gamma_s}{\alpha_s}\mathbb{E}_{\mathcal{F}_{T_s}}\big[\langle\tilde{g}^s - \nabla f_\mu(\tilde{x}^{s-1}), \frac{\alpha_s}{\gamma_s}\sum_{t=1}^{T_s}\theta_t\tilde{x}^s - \big[(1-\alpha_s) + (T_s-1)p_s\big]\tilde{x}^{s-1} - \alpha_s T_s x\rangle\big] \\
&= \mathbb{E}_{\mathcal{F}_{T_s}}\big[\langle\tilde{g}^s - \nabla f_\mu(\tilde{x}^{s-1}), \mathcal{L}_s\tilde{x}^s - \mathcal{R}_s\tilde{x}^{s-1} - \gamma_s T_s x\rangle\big] \\
&= \mathbb{E}_{\mathcal{F}_{T_s}}\big[\langle\tilde{g}^s - \nabla f_\mu(\tilde{x}^{s-1}), \mathcal{L}_s\big(\tilde{x}^s - x\big) - \mathcal{R}_s\big(\tilde{x}^{s-1} - x\big)\rangle\big]. \qquad (34)
\end{aligned}$$

The first equality is indeed the update rule of $\bar{x}_t$, the thrid equality is the definition of $\tilde{x}^s$ and the second last equality comes from the definition of $\mathcal{L}_s$ and $\mathcal{R}_s$.

Then, we set $x = x^*_\mu$ to the inequality above. Based on the assumption **(A2$_\mu$)** and combining the previous inequality with Eq. (33), we have

$$\begin{aligned}\mathcal{L}_s \mathbb{E}_{\mathcal{F}_{T_s}}\big[f_\mu(\tilde{x}^s) - f_\mu(x^*_\mu)\big] \le \mathcal{R}_s \cdot \big[f_\mu(\tilde{x}^{s-1}) - f_\mu(x^*_\mu)\big] + \big(\frac{1}{2}\|x^{s-1} - x^*_\mu\|^2 - \frac{1}{2}\|x^s - x^*_\mu\|^2\big) \\ + T_s \cdot \frac{\gamma_s}{\alpha_s} \cdot \frac{9\alpha_s\gamma_s\mu^2L^2(d+6)^3}{1 - L\alpha_s\gamma_s} + (\mathcal{L}_s + \mathcal{R}_s)Z\|e^s\|.\end{aligned}$$

□

Finally, we derive Theorem 2 directly from Lemma 17. For convenience of the reader, we re-write the theorem here.

Theorem 2. Assume **(A1)** and **(A2$_\mu$)**. If we define $s_0 := \lfloor \log(d+4)n \rfloor + 1$ and set $\{T_s\}$, $\{\gamma_s\}$ and $\{p_s\}$ as

$$T_s = \begin{cases} 2^{s-1}, & s \le s_0 \\ T_{s_0}, & s > s_0 \end{cases}, \quad \gamma_s = \tfrac{1}{12(d+4)L\alpha_s}, \quad p_s = \tfrac{1}{2}, \tag{35}$$

with

$$\alpha_s = \begin{cases} \frac{1}{2}, & s \le s_0 \\ \frac{2}{s - s_0 + 4}, & s > s_0 \end{cases}. \tag{36}$$

If we set

$$\theta_t = \begin{cases} \frac{\gamma_s}{\alpha_s}(\alpha_s + p_s) & 1 \le t \le T_s - 1 \\ \frac{\gamma_s}{\alpha_s} & t = T_s. \end{cases} \tag{37}$$

we obtain

$$\mathbb{E}\big[f_\mu(\tilde{x}^s) - f^*_\mu\big] \le \begin{cases} \dfrac{(d+4)D_0}{2^{s+1}} + \varsigma_1 + \varsigma_2, & 1 \le s \le s_0 \\ \dfrac{16D_0}{n(s - s_0 + 4)^2} + \delta_s \cdot (\varsigma_1 + \varsigma_2), & s > s_0 \end{cases}$$

where $\varsigma_1 = 2\mu^2L(d+4)^2$, $\varsigma_2 = 3Z\sqrt{E}$, $\delta_s = \mathcal{O}(s - s_0)$ and D_0 is defined as

$$D_0 := \frac{2}{(d+4)}[f_\mu(x^0) - f_\mu(x^*_\mu)] + 6L\|x^0 - x^*_\mu\|^2, \tag{38}$$

where x^*_μ is any finite minimizer of f_μ.

Proof of Theorem 2. First, note that, with the parameter choices described in the theorem statement, *the restrictions in Eq.* (25) *and Eq.* (26) *are satisfied*:

$$1 + \tau\gamma_s - L\alpha_s\gamma_s = 1 - \frac{1}{12(d+4)} > 0, \tag{39}$$

$$p_s - \frac{4(d+4)L\alpha_s\gamma_s}{1 + \tau\gamma_s - L\alpha_s\gamma_s} = \frac{1}{2} - \frac{1}{3} \cdot \frac{1}{1 - \frac{1}{12(d+4)}} > 0. \tag{40}$$

We further define

$$w_s := \mathcal{L}_s - \mathcal{R}_{s+1}. \tag{41}$$

As in (Lan et al., 2019), if $1 \le s < s_0$,

$$w_s = \mathcal{L}_s - \mathcal{R}_{s+1} = \frac{\gamma_s}{\alpha_s}\big[1 + (T_s - 1)(\alpha_s + p_s) - (1 - \alpha_s) - (2T_s - 1)p_s\big] = \frac{\gamma_s}{\alpha_s}\big[T_s(\alpha_s - p_s)\big] = 0.$$

Otherwise, if $s \geq s_0$, we have $\frac{\gamma_s}{\alpha_s} = \frac{1}{12(d+4)L\alpha_s^2} = \frac{(s-s_0+4)^2}{48(d+4)L}$ and

$$\begin{aligned} w_s = \mathcal{L}_s - \mathcal{R}_{s+1} &= \frac{\gamma_s}{\alpha_s} - \frac{\gamma_{s+1}}{\alpha_{s+1}}(1-\alpha_{s+1}) + (T_{s_0}-1)\big[\frac{\gamma_s(\alpha_s+p_s)}{\alpha_s} - \frac{\gamma_{s+1}p_{s+1}}{\alpha_{s+1}}\big] \\ &= \frac{(s-s_0+4)^2}{48(d+4)L} - \frac{(s-s_0+5)^2}{48(d+4)L}(1-\frac{2}{s-s_0+5}) \\ &\quad + (T_{s_0}-1)\Big[\frac{(s-s_0+4)^2}{48(d+4)L}\cdot(\frac{2}{s-s_0+4}+\frac{1}{2}) - \frac{(s-s_0+5)^2}{48(d+4)L}\cdot\frac{1}{2}\Big] \\ &= \frac{1}{48(d+4)L} + \frac{T_{s_0}-1}{96(d+4)L}\big[2(s-s_0+4)-1\big] > 0. \end{aligned}$$

Hence, $w_s \geq 0$ for all s. We can therefore use Lemma 17 iteratively as follows,

$$\begin{aligned} &\mathcal{L}_s\mathbb{E}\big[f_\mu(\tilde{x}^s) - f_\mu(x_\mu^*)\big] + \left(\sum_{j=1}^{s-1} w_j\mathbb{E}\big[f_\mu(\tilde{x}^j) - f_\mu(x_\mu^*)\big]\right) \\ &\leq \mathcal{R}_1\cdot\mathbb{E}\big[f_\mu(\tilde{x}^0) - f_\mu(x_\mu^*)\big] + \mathbb{E}\big[\frac{1}{2}\|x^0 - x_\mu^*\|^2 - \frac{1}{2}\|x^s - x_\mu^*\|^2\big] + \sum_{j=1}^{s} T_j\cdot\frac{\gamma_j}{\alpha_j}\cdot\frac{9\alpha_j\gamma_j\mu^2L^2(d+6)^3}{1-L\alpha_j\gamma_j} \\ &\quad + \sum_{j=1}^{s}(\mathcal{L}_j+\mathcal{R}_j)Z\|e^j\| \\ &\leq \frac{1}{6(d+4)L}\big[f_\mu(\tilde{x}^0) - f_\mu(x_\mu^*)\big] + \frac{1}{2}\|x^0 - x_\mu^*\|^2 + \sum_{j=1}^{s} T_j\cdot\frac{\gamma_j}{\alpha_j}\cdot\frac{9\alpha_j\gamma_j\mu^2L^2(d+6)^3}{1-L\alpha_j\gamma_j} \qquad (42) \\ &\quad + \sum_{j=1}^{s}(\mathcal{L}_j+\mathcal{R}_j)Z\|e^j\| \\ &= \frac{1}{12L}D_0 + \sum_{j=1}^{s} T_j\cdot\frac{\gamma_j}{\alpha_j}\cdot\frac{9\alpha_j\gamma_j\mu^2L^2(d+6)^3}{1-L\alpha_j\gamma_j} + \sum_{j=1}^{s}(\mathcal{L}_j+\mathcal{R}_j)Z\|e^j\| \\ &\leq \frac{1}{12L}D_0 + \sum_{j=1}^{s} T_j\cdot\frac{\gamma_j}{\alpha_j}\cdot\frac{3\mu^2L(d+6)^3}{4(d+4)} + \sum_{j=1}^{s}(\mathcal{L}_j+\mathcal{R}_j)Z\|e^j\| \\ &\leq \frac{1}{12L}D_0 + \sum_{j=1}^{s} T_j\cdot\frac{\gamma_j}{\alpha_j}\cdot\mu^2L(d+4)^2 + \sum_{j=1}^{s}(\mathcal{L}_j+\mathcal{R}_j)Z\|e^j\| \\ &\leq \frac{1}{12L}D_0 + \sum_{j=1}^{s} T_j\cdot\frac{\gamma_j}{\alpha_j}\cdot\mu^2L(d+4)^2 + \sum_{j=1}^{s}(\mathcal{L}_j+\mathcal{R}_j)Z\sqrt{E}. \qquad (43) \end{aligned}$$

The second last equality holds when $\alpha_j\gamma_j = \frac{1}{12(d+4)L}$ and $x = x_\mu^*$, the optimal solution for f_μ. We proceed with two cases:

Case I: If $s \leq s_0$, $\mathcal{L}_s = \frac{2^{s+1}}{12(d+4)L}$, $\mathcal{R}_s = \frac{2^s}{12(d+4)L} = \frac{\mathcal{L}_s}{2}$, $\frac{\gamma_s}{\alpha_s} = \frac{1}{3(d+4)L}$, $T_s = 2^{s-1}$. Hence, we have

$$\mathbb{E}\big[f_\mu(\tilde{x}^s) - f_\mu(x_\mu^*)\big] \leq \frac{1}{2^{s+1}}(d+4)D_0 + 2\mu^2L(d+4)^2 + 3Z\sqrt{E}, \qquad 1 \leq s \leq s_0. \qquad (44)$$

Case II: If $s > s_0$, we have

$$\begin{aligned} \mathcal{L}_s &= \frac{1}{12(d+4)\alpha_s^2}\big[(T_s-1)\alpha_s + \frac{1}{2}(T_s+1)\big] = \frac{(s-s_0+4)^2}{48(d+4)L}\cdot\big[(T_{s_0}-1)\alpha_s + \frac{1}{2}(T_{s_0}+1)\big] \\ &\geq \frac{(s-s_0+4)^2}{96(d+4)L}\cdot(T_{s_0}+1) \geq \frac{n\cdot(s-s_0+4)^2}{192L}, \end{aligned}$$

where the last inequality holds since $T_{s_0} = 2^{\lfloor \log_2 [(d+4)n] \rfloor} \geq \frac{(d+4)n}{2}$, i.e. $2^{s_0} \geq (d+4)n$. Hence, based on $\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$ and Eq. (44), Eq. (43) implies

$$\mathbb{E}\big[f_\mu(\tilde{x}^s) - f_\mu(x_\mu^*)\big] \leq \frac{16D_0}{n(s-s_0+4)^2} + \mathcal{O}(s-s_0)\cdot \mu^2 L(d+4)^2 + \mathcal{O}(s-s_0)\cdot Z\sqrt{E}. \tag{45}$$

□

We conclude by deriving the final complexity result, stated in the main paper.

Proof of Corollary 3. We pick up from the proof presented in the main paper, which we summarize in the next lines. Note that the analysis we performed in the last pages is based on f_μ rather than f. Hence, we first need to ensure that the error between these two functions is sufficiently small. We can bound $f_\mu(\tilde{x}^s) - f_\mu(x_\mu^*)$ from $f(\tilde{x}^s) - f(x^*)$ as follows:

$$\begin{aligned} f_\mu(\tilde{x}^s) - f_\mu(x_\mu^*) &= f_\mu(\tilde{x}^s) - f(\tilde{x}^s) + f(\tilde{x}^s) - f_\mu(x_\mu^*) + f_\mu(x^*) - f_\mu(x^*) + f(x^*) - f(x^*) \\ &\geq -\frac{\mu^2 Ld}{2} + f(\tilde{x}^s) - f_\mu(x_\mu^*) + f_\mu(x^*) - \frac{\mu^2 Ld}{2} - f(x^*) \\ &\geq f(\tilde{x}^s) - f(x^*) - \mu^2 Ld. \end{aligned}$$

The first inequality comes from Lemma 11 and the second inequality comes from the definition of x_μ.

We want the error term $\mu^2 Ld$ we just derived to be small, say $\leq \frac{\epsilon}{4}$, i.e. $\mu = \mathcal{O}\big(\sqrt{\frac{\epsilon}{4Ld}}\big)$. From this, we get an upper bound on μ (*choosing μ small does not affect the convergence rate*). Next, we bound in the same way the additional (non-vanishing) error terms in Eq. (44) and Eq. (45). This requires $\mu = \mathcal{O}\big(\frac{\epsilon^{1/2}}{L^{1/2}d}\big)$, $\mu = \mathcal{O}\big(\frac{\epsilon}{ZLd^{3/2}}\big)$ and $\nu = \mathcal{O}\big(\frac{\epsilon}{ZLd^{1/2}}\big)$ for Eq. (44) while $\mu = \mathcal{O}\big(\frac{n^{1/4}\epsilon^{3/4}}{L^{1/2}dD_0^{1/4}}\big)$, $\mu = \mathcal{O}\big(\frac{n^{1/2}\epsilon^{3/2}}{ZLd^{3/2}D_0^{1/2}}\big)$ and $\nu = \mathcal{O}\big(\frac{n^{1/2}\epsilon^{3/2}}{ZLd^{1/2}D_0^{1/2}}\big)$ for Eq. (45) to ensure ϵ-optimality, $\frac{\epsilon}{4}$ more specifically. Therefore, if we bound the term which contains D_0 by $\frac{\epsilon}{2}$, $f(\tilde{x}^s) - f(x^*)$ would achieve ϵ-optimality in expectation. This is what we do next (following the proof in (Lan, 2012)), for the two cases in Theorem 2.

If $n \geq \frac{D_0}{\epsilon}$, i.e. in the region of relatively low accuracy and/or large number of components, we have

$$\frac{(d+4)D_0}{2^{s_0+1}} \leq \frac{D_0}{2n} \leq \frac{\epsilon}{2} \Rightarrow \log\frac{(d+4)D_0}{\epsilon} \leq s_0.$$

Therefore, the number of epochs is at most s_0 for the first term in Eq. (44) to achieve $\frac{\epsilon}{2}$ optimality inside Case I. Hence, the total number of function queries is bounded by

$$dnS_l + \sum_{s=1}^{S_l} T_s = \mathcal{O}\Bigg\{\min\bigg(dn\log\frac{(d+4)D_0}{\epsilon}, dn\log(dn), dn\bigg)\Bigg\} = \mathcal{O}\Bigg\{\min\bigg(dn\log\frac{dD_0}{\epsilon}, dn\bigg)\Bigg\} = \mathcal{O}\left\{dn\log\frac{dD_0}{\epsilon}\right\}.$$

If instead $n < \frac{D_0}{\epsilon}$, at epoch $S_h = \left\lceil \sqrt{\frac{32D_0}{n\epsilon}} + s_0 - 4 \right\rceil$ (ensuring the first term in Eq. (45) to be not bigger than $\frac{\epsilon}{2}$), we can achieve ϵ optimality. Hence, the total number of function queries is

$$dns_0 + \sum_{s=1}^{s_0} T_s + (T_{s_0} + dn)(S_h - s_0) \leq \sum_{s=1}^{s_0} T_s + (T_{s_0} + dn)S_h = \mathcal{O}\bigg\{d\sqrt{\frac{nD_0}{\epsilon}} + dn\log(dn)\bigg\}.$$

□

B.2. Proof of Theorem 4

In this section, we assume f to be strongly convex. Hence, f_μ is also strongly convex by Lemma 10.

(A3) $f = \frac{1}{n}\sum_{i=1}^n f_i$ is τ-strongly convex. That is, for all $x, y \in \mathbb{R}^d$, $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\tau}{2}\|y - x\|^2$.

We rewrite below Theorem 4, for convenience of the reader:

Theorem 4. Assume **(A1)** and **(A3)**. Let us denote $s_0 := \lfloor \log(d+4)n \rfloor + 1$ and assume that the weights $\{\theta_t\}$ are set to Eq. (10) if $1 \le s \le s_0$. Otherwise, they are set to

$$\theta_t = \begin{cases} \Gamma_{t-1} - (1-\alpha_s - p_s)\Gamma_t, & 1 \le t \le T_s - 1, \\ \Gamma_{t-1}, & t = T_s, \end{cases} \tag{46}$$

where $\Gamma_t = \left(1 + \frac{\tau\gamma_s}{2}\right)^t$. If the parameters $\{T_s\}$, $\{\gamma_s\}$ and $\{p_s\}$ are set to Eq. (8) with

$$\alpha_s = \begin{cases} \frac{1}{2}, & s \le s_0, \\ \min\{\sqrt{\frac{n\tau}{24L}}, \frac{1}{2}\}, & s > s_0, \end{cases} \tag{47}$$

we obtain

$$\mathbb{E}\left[f_\mu(\tilde{x}^s) - f_\mu^*\right] \le \begin{cases} \dfrac{1}{2^{s+1}}(d+4)D_0 + \varsigma_1 + \varsigma_2, & 1 \le s \le s_0 \\ (4/5)^{s-s_0}\dfrac{D_0}{n} + \varsigma_1 + \varsigma_2, & s > s_0 \text{ and } n \ge \frac{6L}{\tau} \\ \left(1 + \frac{1}{4}\sqrt{\frac{n\tau}{6L}}\right)^{-(s-s_0)}\dfrac{D_0}{n} + \left(\sqrt{\frac{6L}{n\tau}} + 1\right)\varsigma_1 + \varsigma_2 & s > s_0 \text{ and } n < \frac{6L}{\tau} \end{cases}$$

where $\varsigma_1 = 12\mu^2 L(d+4)^2$, $\varsigma_2 = 5E/\tau$ and D_0 is defined as in Eq. (11).

Remark. *Compared with smooth convex case, we can drop the assumption **(A2$_\mu$)**.*

We start with the following result.

Lemma 18. *Assume **(A1)**, **(A3)**. Under the choice of parameters from Theorem 4, for any $0 < c \le 1$,*

$$\begin{aligned} &\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\frac{\gamma_s}{\alpha_s}\left[f_\mu(\bar{x}_t) - f_\mu^*\right] + \left(1 + (1-c)\tau\gamma_s\right)\cdot\frac{1}{2}\|x_t - x_\mu^*\|^2\right] \\ &\le \frac{\gamma_s}{\alpha_s}(1-\alpha_s-p_s)\left[f_\mu(\bar{x}_{t-1}) - f_\mu^*\right] + \frac{\gamma_s p_s}{\alpha_s}\left[f_\mu(\tilde{x}) - f_\mu^*\right] + \frac{1}{2}\|x_{t-1} - x_\mu^*\|^2 + \frac{\gamma_s}{\alpha_s}\cdot\mu^2 L(d+4)^2 + \frac{\gamma_s}{2c\tau}\cdot E, \end{aligned} \tag{48}$$

where E is defined as in Lemma 17.

Proof of Lemma 18. First, note that, with the parameter choices described in the Theorem 4, *the restrictions in Eq.* (25) *and Eq.* (26) *are satisfied*. Hence, Eq. (27) becomes, when setting $x = x_\mu^*$,

$$\begin{aligned} &\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\frac{\gamma_s}{\alpha_s}\left[f_\mu(\bar{x}_t) - f_\mu^*\right] + \frac{(1+\tau\gamma_s)}{2}\|x_t - x_\mu^*\|^2\right] \\ &\le \frac{\gamma_s}{\alpha_s}(1-\alpha_s-p_s)\left[f_\mu(\bar{x}_{t-1}) - f_\mu^*\right] + \frac{\gamma_s p_s}{\alpha_s}\left[f_\mu(\tilde{x}) - f_\mu^*\right] + \frac{1}{2}\|x_{t-1} - x_\mu^*\|^2 + \frac{\gamma_s}{\alpha_s}\cdot\frac{9\alpha_s\gamma_s\mu^2 L^2(d+6)^3}{1 + \tau\gamma_s - L\alpha_s\gamma_s} \\ &\quad - \gamma_s\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\langle \tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x_\mu^*\rangle\right] \\ &= \frac{\gamma_s}{\alpha_s}(1-\alpha_s-p_s)\left[f_\mu(\bar{x}_{t-1}) - f_\mu^*\right] + \frac{\gamma_s p_s}{\alpha_s}\left[f_\mu(\tilde{x}) - f_\mu^*\right] + \frac{1}{2}\|x_{t-1} - x_\mu^*\|^2 + \frac{\gamma_s}{\alpha_s}\cdot\frac{3\mu^2 L(d+6)^3}{4(1 + \tau\gamma_s - \frac{1}{12})(d+4)} \\ &\quad - \gamma_s\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\langle \tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x_\mu^*\rangle\right] \\ &\le \frac{\gamma_s}{\alpha_s}(1-\alpha_s-p_s)\left[f_\mu(\bar{x}_{t-1}) - f_\mu^*\right] + \frac{\gamma_s p_s}{\alpha_s}\left[f_\mu(\tilde{x}) - f_\mu^*\right] + \frac{1}{2}\|x_{t-1} - x_\mu^*\|^2 + \frac{\gamma_s}{\alpha_s}\cdot\mu^2 L(d+4)^2 \\ &\quad - \gamma_s\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\langle \tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x_\mu^*\rangle\right]. \end{aligned}$$

The equality above holds since α_s and γ_s are defined as in Theorem 4 since $\alpha_s\gamma_s = \frac{1}{12(d+4)L}$.
Moreover, for any $0 < c \leq 1$, we have

$$-\gamma_s\langle \tilde{g} - \nabla f_\mu(\tilde{x}), x_t - x_\mu^*\rangle - \frac{c\tau\gamma_s}{2}\|x_t - x_\mu^*\|^2 \leq \frac{\gamma_s}{2c\tau}\|\tilde{g} - \nabla f_\mu(\tilde{x})\|^2,$$

since $b\langle u, v\rangle - \frac{a}{2}\|v\|^2 \leq \frac{b^2}{2a}\|u\|^2$ when $a > 0$. Hence, plugging this in,

$$\begin{aligned}
&\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\frac{\gamma_s}{\alpha_s}\big[f_\mu(\bar{x}_t) - f_\mu^*\big] + \big(1 + (1-c)\tau\gamma_s\big)\cdot\frac{1}{2}\|x_t - x_\mu^*\|^2\right]\\
\leq\ &\frac{\gamma_s}{\alpha_s}(1 - \alpha_s - p_s)\big[f_\mu(\bar{x}_{t-1}) - f_\mu^*\big] + \frac{\gamma_s p_s}{\alpha_s}\big[f_\mu(\tilde{x}) - f_\mu^*\big] + \frac{1}{2}\|x_{t-1} - x_\mu^*\|^2 + \frac{\gamma_s}{\alpha_s}\cdot\mu^2 L(d+4)^2\\
&+ \frac{\gamma_s}{2c\tau}\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\big[\|\tilde{g} - \nabla f_\mu(\tilde{x})\|^2\big]\\
\leq\ &\frac{\gamma_s}{\alpha_s}(1 - \alpha_s - p_s)\big[f_\mu(\bar{x}_{t-1}) - f_\mu^*\big] + \frac{\gamma_s p_s}{\alpha_s}\big[f_\mu(\tilde{x}) - f_\mu^*\big] + \frac{1}{2}\|x_{t-1} - x_\mu^*\|^2 + \frac{\gamma_s}{\alpha_s}\cdot\mu^2 L(d+4)^2 + \frac{\gamma_s}{2c\tau}\cdot E.
\end{aligned}$$

□

We divide the proof of Theorem 4 into three cases, corresponding to the three lemmas below.

Lemma 19. *Assume **(A1)**, **(A3)**. Under the choice of parameters from Theorem 4, if $s \leq s_0$, then for any $x \in \mathbb{R}^d$,*

$$\mathbb{E}\big[f_\mu(\tilde{x}^s) - f_\mu^*\big] \leq \frac{1}{2^{s+1}}(d+4)D_0 + 2\mu^2 L(d+4)^2 + \frac{E}{2\tau},$$

where D_0 is defined in Eq. (11).

Proof of Lemma 19. For this case, $\alpha_s = p_s = \frac{1}{2}$, $\gamma_s = \frac{1}{6(d+4)L}$, $T_s = 2^{s-1}$. Starting from Lemma 18, if we set $c = 1$ in Eq. (48) and sum it up from $t = 1$ to T_s, we have

$$\begin{aligned}
&\sum_{t=1}^{T_s}\frac{\gamma_s}{\alpha_s}\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\big[f_\mu(\bar{x}_t) - f_\mu^*\big] + \frac{1}{2}\mathbb{E}\big[\|x_{T_s} - x_\mu^*\|^2\big]\\
\leq\ &\frac{\gamma_s}{2\alpha_s}\cdot T_s\big[f_\mu(\tilde{x}) - f_\mu^*\big] + \frac{1}{2}\|x_0 - x_\mu^*\|^2 + \frac{\gamma_s}{\alpha_s}\cdot T_s\cdot\mu^2 L(d+4)^2 + T_s\cdot\frac{\gamma_s}{2\tau}\cdot E.
\end{aligned}$$

Thanks to convexity of f_μ, we have $f\left(\frac{1}{T_s}\sum_{t=1}^{T_s}\bar{x}_t\right) \leq \frac{1}{T_s}\sum_{t=1}^{T^s} f(\bar{x}_t)$. Hence, the last inequality implies

$$\begin{aligned}
&\frac{1}{3(d+4)L}\cdot T_s\cdot\mathbb{E}_{\mathcal{F}_{T_s}}\big[f_\mu(\tilde{x}^s) - f_\mu^*\big] + \frac{1}{2}\mathbb{E}_{\mathcal{F}_{T_s}}\big[\|x^s - x_\mu^*\|^2\big]\\
\leq\ &\frac{1}{6(d+4)L}\cdot T_s\big[f_\mu(\tilde{x}^{s-1}) - f_\mu^*\big] + \frac{1}{2}\|x^{s-1} - x_\mu^*\|^2 + \frac{1}{3(d+4)L}\cdot T_s\cdot\mu^2 L(d+4)^2 + T_s\cdot\frac{1}{12(d+4)L\tau}\cdot E\\
=\ &\frac{1}{3(d+4)L}\cdot T_{s-1}\big[f_\mu(\tilde{x}^{s-1}) - f_\mu^*\big] + \frac{1}{2}\|x^{s-1} - x_\mu^*\|^2 + \frac{1}{3(d+4)L}\cdot T_s\cdot\mu^2 L(d+4)^2 + T_s\cdot\frac{1}{12(d+4)L\tau}\cdot E,
\end{aligned}$$

where $x_{T_s} = x^s$, $x_0 = x^{s-1}$, $\tilde{x} = \tilde{x}^{s-1}$. Applying the last inequality iteratively, we obtain

$$\begin{aligned}
&\frac{1}{3(d+4)L}\cdot T_s\cdot\mathbb{E}\big[f_\mu(\tilde{x}^s) - f_\mu^*\big] + \frac{1}{2}\mathbb{E}\big[\|x^s - x_\mu^*\|^2\big]\\
\leq\ &\frac{1}{3(d+4)L}\cdot T_0\big[f_\mu(\tilde{x}^0) - f_\mu^*\big] + \frac{1}{2}\|x^0 - x_\mu^*\|^2 + \frac{1}{3(d+4)L}\sum_{j=1}^{s}T_j\cdot\mu^2 L(d+4)^2 + \sum_{j=1}^{s}T_j\cdot\frac{1}{12(d+4)L\tau}\cdot E,
\end{aligned}$$

where $T_0 = \frac{1}{2}$ is in accordance with the definition of $T_s = 2^{s-1},\ 0 < s \leq s_0$. Finally, we obtain

$$\mathbb{E}\big[f_\mu(\tilde{x}^s) - f_\mu^*\big] + \frac{\alpha_s}{\gamma_s T_s}\cdot\frac{1}{2}\mathbb{E}\big[\|x^s - x_\mu^*\|^2\big]$$

$$
\begin{aligned}
&= \mathbb{E}\big[f_\mu(\tilde{x}^s) - f_\mu^*\big] + \frac{3(d+4)L}{T_s} \cdot \frac{1}{2}\mathbb{E}\big[\|x^s - x_\mu^*\|^2\big] \\
&\le \frac{1}{2^s}\big[f_\mu(\tilde{x}^0) - f_\mu^* + 3(d+4)L\|x^0 - x_\mu^*\|^2\big] + \frac{1}{2^{s-1}}\sum_{j=1}^{s} T_j \cdot \mu^2 L(d+4)^2 + \frac{1}{2^{s-1}}\sum_{j=1}^{s} T_j \cdot \frac{1}{4\tau} \cdot E \\
&\le \frac{1}{2^{s+1}}(d+4)D_0 + 2\mu^2 L(d+4)^2 + \frac{E}{2\tau}. \qquad (49)
\end{aligned}
$$

We conclude the proof by observing that $\frac{1}{2^{s-1}}\sum_{j=1}^{s} T_j \le 2$ when $s \le s_0$. □

Lemma 20. *Assume (A1), (A3). Under the choice of parameters from Theorem 4, if $s \ge s_0$ and $n \ge \frac{6L}{\tau}$, then for any $x \in \mathbb{R}^d$*

$$
\mathbb{E}\big[f_\mu(\tilde{x}^s) - f_\mu^*\big] \le \left(\frac{4}{5}\right)^{s-s_0}\frac{D_0}{n} + 12\mu^2 L(d+4)^2 + \frac{5E}{\tau}.
$$

Proof of Lemma 20. For this case, $\alpha_s = \alpha = p_s = \frac{1}{2}$, $\gamma_s = \gamma = \frac{1}{6(d+4)L}$, $T_s = 2^{s_0-1}$ when $s \ge s_0$. Thanks to Lemma 18, if we set $c = \frac{1}{2}$ in Eq. (48), we have

$$
\begin{aligned}
&\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\frac{\gamma}{\alpha}\big[f_\mu(\bar{x}_t) - f_\mu^*\big] + \big(1 + \frac{\tau\gamma}{2}\big)\cdot\frac{1}{2}\|x_t - x_\mu^*\|^2\right] \\
&\qquad\le \frac{\gamma}{2\alpha}\big[f_\mu(\tilde{x}) - f_\mu^*\big] + \frac{1}{2}\|x_{t-1} - x_\mu^*\|^2 + \frac{\gamma}{\alpha}\cdot\mu^2 L(d+4)^2 + \frac{\gamma}{\tau}\cdot E.
\end{aligned}
$$

Multiplying both sides by $\Gamma_{t-1} = (1 + \frac{\tau\gamma}{2})^{t-1}$, we obtain

$$
\begin{aligned}
&\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\frac{\gamma}{\alpha}\Gamma_{t-1}\big[f_\mu(\bar{x}_t) - f_\mu^*\big] + \frac{\Gamma_t}{2}\|x_t - x_\mu^*\|^2\right] \\
&\le \frac{\gamma}{2\alpha}\Gamma_{t-1}\big[f_\mu(\tilde{x}) - f_\mu^*\big] + \frac{\Gamma_{t-1}}{2}\|x_{t-1} - x_\mu^*\|^2 + \frac{\gamma}{\alpha}\Gamma_{t-1}\cdot\mu^2 L(d+4)^2 + \frac{\gamma}{\tau}\Gamma_{t-1}\cdot E.
\end{aligned}
$$

Since $\theta_t = \Gamma_{t-1}$ as defined in Eq. (12), the last inequality can be rewritten as

$$
\begin{aligned}
&\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\frac{\gamma}{\alpha}\theta_t\big[f_\mu(\bar{x}_t) - f_\mu^*\big] + \frac{\Gamma_t}{2}\|x_t - x_\mu^*\|^2\right] \\
&\qquad\le \frac{\gamma}{2\alpha}\theta_t\big[f_\mu(\tilde{x}) - f_\mu^*\big] + \frac{\Gamma_{t-1}}{2}\|x_{t-1} - x_\mu^*\|^2 + \frac{\gamma}{\alpha}\theta_t\cdot\mu^2 L(d+4)^2 + \frac{\gamma}{\tau}\theta_t\cdot E.
\end{aligned}
$$

Summing up the inequality above from $t = 1$ to T_s, we obtain

$$
\begin{aligned}
&\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\theta_t\mathbb{E}_{\mathcal{F}_t}\big[f_\mu(\bar{x}_t) - f_\mu^*\big] + \frac{\Gamma_{T_s}}{2}\mathbb{E}_{\mathcal{F}_{T_s}}\|x_{T_s} - x_\mu^*\|^2 \\
&\le \frac{\gamma}{2\alpha}\sum_{t=1}^{T_s}\theta_t\mathbb{E}_{\mathcal{F}_{T_s}}\big[f_\mu(\tilde{x}) - f_\mu^*\big] + \frac{1}{2}\mathbb{E}_{\mathcal{F}_{T_s}}\big[\|x_0 - x_\mu^*\|^2\big] + \frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\theta_t\cdot\mu^2 L(d+4)^2 + \frac{\gamma}{\tau}\sum_{t=1}^{T_s}\theta_t\cdot E,
\end{aligned}
$$

and then

$$
\begin{aligned}
&\frac{5}{4}\left[\frac{\gamma}{2\alpha}\sum_{t=1}^{T_s}\theta_t\mathbb{E}_{\mathcal{F}_t}\big[f_\mu(\bar{x}_t) - f_\mu^*\big] + \frac{1}{2}\mathbb{E}_{\mathcal{F}_{T_s}}\|x_{T_s} - x_\mu^*\|^2\right] \\
&\le \frac{\gamma}{2\alpha}\sum_{t=1}^{T_s}\theta_t\mathbb{E}_{\mathcal{F}_{T_s}}\big[f_\mu(\tilde{x}) - f_\mu^*\big] + \frac{1}{2}\mathbb{E}_{\mathcal{F}_{T_s}}\big[\|x_0 - x_\mu^*\|^2\big] + \frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\theta_t\cdot\mu^2 L(d+4)^2 + \frac{\gamma}{\tau}\sum_{t=1}^{T_s}\theta_t\cdot E, \qquad (50)
\end{aligned}
$$

The last inequality is based on the fact that, for $s \geq s_0$, $\frac{(d+4)n}{2} \leq T_s = T_{s_0} \leq (d+4)n$, we have

$$\begin{aligned}\Gamma_{T_s} &= \big(1+\frac{\tau\gamma}{2}\big)^{T_s} = \big(1+\frac{\tau\gamma}{2}\big)^{T_{s_0}} \geq 1+\frac{\tau\gamma}{2}\cdot T_{s_0} \geq 1+\frac{\tau\gamma}{2}\cdot\frac{(d+4)n}{2}\\ &= 1+\frac{\tau}{12(d+4)L}\cdot\frac{(d+4)n}{2} = 1+\frac{\tau n}{24L} \geq \frac{5}{4},\end{aligned}$$

where the last step holds under $n \geq \frac{6L}{\tau}$. Since $\tilde{x}^s = \sum_{t=1}^{T_s}(\theta_t \bar{x}_t)/\sum_{t=1}^{T_s}\theta_t$, $\tilde{x} = \tilde{x}^{s-1}$, $x_0 = x^{s-1}$, $x_{T_s} = x^s$ in the epoch s and the convexity of f_μ, Eq. (50) implies

$$\begin{aligned}&\frac{5}{4}\Big[\frac{\gamma}{2\alpha}\mathbb{E}_{\mathcal{F}_{T_s}}\big[f_\mu(\tilde{x}^s)-f_\mu^*\big] + \frac{1}{2\sum_{t=1}^{T_s}\theta_t}\mathbb{E}_{\mathcal{F}_{T_s}}\|x^s-x_\mu^*\|^2\Big]\\ \leq\ &\frac{\gamma}{2\alpha}\mathbb{E}_{\mathcal{F}_{T_s}}\big[f_\mu(\tilde{x}^{s-1})-f_\mu^*\big] + \frac{1}{2\sum_{t=1}^{T_s}\theta_t}\mathbb{E}_{\mathcal{F}_{T_s}}\big[\|x^{s-1}-x_\mu^*\|^2\big] + \frac{\gamma}{\alpha}\cdot\mu^2L(d+4)^2 + \frac{\gamma}{\tau}\cdot E.\end{aligned}$$

Applying it recursively for $s \geq s_0$, we obtain

$$\begin{aligned}&\mathbb{E}\big[f_\mu(\tilde{x}^s)-f_\mu^*\big] + \frac{2\alpha}{\gamma\sum_{t=1}^{T_s}\theta_t}\cdot\frac{1}{2}\mathbb{E}\big[\|x^s-x_\mu^*\|^2\big]\\ \leq\ &\big(4/5\big)^{s-s_0}\Big[\mathbb{E}\big[f_\mu(\tilde{x}^{s_0})-f_\mu^*\big] + \frac{2\alpha}{\gamma\sum_{t=1}^{T_s}\theta_t}\cdot\frac{1}{2}\mathbb{E}\big[\|x^{s_0}-x_\mu^*\|^2 + \sum_{j=s_0+1}^{s}\big(4/5\big)^{s+1-j}\Big[2\mu^2L(d+4)^2 + \frac{2\alpha}{\tau}\cdot E\Big]\\ \leq\ &\big(4/5\big)^{s-s_0}\Big[\mathbb{E}\big[f_\mu(\tilde{x}^{s_0})-f_\mu^*\big] + \frac{2\alpha}{\gamma T_{s_0}}\cdot\frac{1}{2}\mathbb{E}\big[\|x^{s_0}-x_\mu^*\|^2\big]\Big] + 8\mu^2L(d+4)^2 + \frac{4}{\tau}\cdot E,\end{aligned}$$

where the last inequality holds because $\sum_{j=s_0+1}^{s}\big(\frac{4}{5}\big)^{s+1-j} \leq \frac{4}{5}\cdot\frac{1}{1-\frac{4}{5}} = 4$, $\sum_{t=1}^{T_s}\theta_t \geq T_s = T_{s_0}$ and $2\alpha \leq 1$. Finally,

$$\begin{aligned}&\mathbb{E}\big[f_\mu(\tilde{x}^s)-f_\mu^*\big] + \frac{2\alpha}{\gamma\sum_{t=1}^{T_s}\theta_t}\cdot\frac{1}{2}\mathbb{E}\big[\|x^s-x_\mu^*\|^2\big]\\ \leq\ &\big(4/5\big)^{s-s_0}\Big[\mathbb{E}\big[f_\mu(\tilde{x}^{s_0})-f_\mu^*\big] + \frac{2\alpha}{\gamma T_{s_0}}\cdot\frac{1}{2}\mathbb{E}\big[\|x^{s_0}-x_\mu^*\|^2\big]\Big] + 8\mu^2L(d+4)^2 + \frac{4}{\tau}\cdot E\\ \leq\ &\big(4/5\big)^{s-s_0}2\Big[\mathbb{E}\big[f_\mu(\tilde{x}^{s_0})-f_\mu^*\big] + \frac{\alpha}{\gamma T_{s_0}}\cdot\frac{1}{2}\mathbb{E}\big[\|x^{s_0}-x_\mu^*\|^2\big]\Big] + 8\mu^2L(d+4)^2 + \frac{4}{\tau}\cdot E\\ \leq\ &\big(4/5\big)^{s-s_0}2\cdot\big[\frac{1}{2^{s_0+1}}(d+4)D_0 + 2\mu^2L(d+4)^2 + \frac{E}{2\tau}\big] + 8\mu^2L(d+4)^2 + \frac{4}{\tau}\cdot E\\ \leq\ &\big(4/5\big)^{s-s_0}\cdot\frac{(d+4)D_0}{2^{s_0}} + 12\mu^2L(d+4)^2 + \frac{5E}{\tau}\\ =\ &\big(4/5\big)^{s-s_0}\frac{(d+4)D_0}{2T_{s_0}} + 12\mu^2L(d+4)^2 + \frac{5E}{\tau}\\ \leq\ &\big(4/5\big)^{s-s_0}\frac{D_0}{n} + 12\mu^2L(d+4)^2 + \frac{5E}{\tau},\end{aligned}$$

where the third inequality comes from Eq. (49) and the last inquality relies on $T_{s_0} \geq \frac{(d+4)n}{2}$. □

Lemma 21. *Assume (A1), (A3). Under the choice of parameters from Theorem 4, if* $s \geq s_0$ *and* $n < \frac{6L}{\tau}$*, then for any* $x \in \mathbb{R}^d$*,*

$$\mathbb{E}\big[f_\mu(\tilde{x}^s)-f_\mu^*\big] \leq \left(1+\frac{1}{4}\sqrt{\frac{n\tau}{6L}}\right)^{-(s-s_0)}\frac{D_0}{n} + \left(8\sqrt{\frac{6L}{n\tau}}+4\right)\mu^2L(d+4)^2 + \frac{5E}{\tau}.$$

Proof of Lemma 21. For this case, $\alpha_s = \alpha = \sqrt{\frac{n\tau}{24L}}$, $p_s = p = \frac{1}{2}$, $\gamma_s = \gamma = \frac{1}{(d+4)\sqrt{6nL\tau}}$, $T_s = T_{s_0} = 2^{s_0-1}$ when $s \geq s_0$. Based on Lemma 18, if we set $c = \frac{1}{2}$ in Eq. (48), we have

$$\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\Big[\frac{\gamma}{\alpha}\big[f_\mu(\bar{x}_t)-f_\mu^*\big] + \big(1+\frac{\tau\gamma}{2}\big)\cdot\frac{1}{2}\|x_t-x_\mu^*\|^2\Big]$$

$$\leq \frac{\gamma}{\alpha}(1-\alpha-p)\big[f_\mu(\bar{x}_{t-1})-f_\mu^*\big]+\frac{\gamma}{2\alpha}\big[f_\mu(\tilde{x})-f_\mu^*\big]+\frac{1}{2}\|x_{t-1}-x_\mu^*\|^2+\frac{\gamma}{\alpha}\cdot\mu^2L(d+4)^2+\frac{\gamma}{\tau}\cdot E.$$

Multiplying both sides by $\Gamma_{t-1}=(1+\frac{\tau\gamma}{2})^{t-1}$, we obtain

$$\begin{aligned}
&\mathbb{E}_{u_t,i_t|\mathcal{F}_{t-1}}\left[\frac{\gamma}{\alpha}\Gamma_{t-1}\big[f_\mu(\bar{x}_t)-f_\mu^*\big]+\frac{\Gamma_t}{2}\|x_t-x_\mu^*\|^2\right]\\
&\leq \frac{\Gamma_{t-1}\gamma}{\alpha}(1-\alpha-p)\big[f_\mu(\bar{x}_{t-1})-f_\mu^*\big]+\frac{\Gamma_{t-1}\gamma p}{\alpha}\big[f_\mu(\tilde{x})-f_\mu^*\big]\\
&\quad+\frac{\Gamma_{t-1}}{2}\|x_{t-1}-x_\mu^*\|^2+\frac{\gamma}{\alpha}\Gamma_{t-1}\cdot\mu^2L(d+4)^2+\frac{\gamma}{\tau}\Gamma_{t-1}\cdot E.
\end{aligned}$$

Summing up the inequality above from $t=1$ to T_s, we obtain

$$\begin{aligned}
&\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\theta_t\mathbb{E}_{\mathcal{F}_t}\left[f_\mu(\bar{x}_t)-f_\mu^*\right]+\frac{\Gamma_{T_s}}{2}\mathbb{E}_{\mathcal{F}_t}\|x_{T_s}-x_\mu^*\|^2\\
&\leq \frac{\gamma}{\alpha}\big[1-\alpha-p+p\sum_{t=1}^{T_s}\Gamma_{t-1}\big]\left[f_\mu(\tilde{x})-f_\mu^*\right]+\frac{1}{2}\|x_0-x_\mu^*\|^2+\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\Gamma_{t-1}\cdot\mu^2L(d+4)^2+\frac{\gamma}{\tau}\sum_{t=1}^{T_s}\Gamma_{t-1}\cdot E.
\end{aligned}$$

Since $\tilde{x}^s=\sum_{t=1}^{T_s}(\theta_t\bar{x}_t)/\sum_{t=1}^{T_s}\theta_t$, $\tilde{x}=\tilde{x}^{s-1}$, $x_0=x^{s-1}$, $x_{T_s}=x^s$ in the epoch s, and thanks to the convexity of f_μ, the last inequality implies, for $s>s_0$:

$$\begin{aligned}
&\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\theta_t\mathbb{E}_{\mathcal{F}_{T_s}}\left[f_\mu(\tilde{x}^s)-f_\mu^*\right]+\frac{\Gamma_{T_{s_0}}}{2}\mathbb{E}_{\mathcal{F}_{T_s}}\|x^s-x_\mu^*\|^2\\
&\leq \frac{\gamma}{\alpha}\big[1-\alpha-p+p\sum_{t=1}^{T_s}\Gamma_{t-1}\big]\left[f_\mu(\tilde{x}^{s-1})-f_\mu^*\right]+\frac{1}{2}\|x^{s-1}-x_\mu^*\|^2+\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\Gamma_{t-1}\cdot\mu^2L(d+4)^2+\frac{\gamma}{\tau}\sum_{t=1}^{T_s}\Gamma_{t-1}\cdot E.
\end{aligned} \tag{51}$$

Moreover, we have

$$\begin{aligned}
\sum_{t=1}^{T_{s_0}}\theta_t&=\Gamma_{T_{s_0}-1}+\sum_{t=1}^{T_{s_0}-1}\big(\Gamma_{t-1}-(1-\alpha-p)\Gamma_t\big)\\
&=\Gamma_{T_{s_0}}(1-\alpha-p)+\sum_{t=1}^{T_{s_0}}\big(\Gamma_{t-1}-(1-\alpha-p)\Gamma_t\big)\\
&=\Gamma_{T_{s_0}}(1-\alpha-p)+\big[1-(1-\alpha-p)(1+\frac{\tau\gamma}{2})\big]\sum_{t=1}^{T_{s_0}}\Gamma_{t-1}.
\end{aligned}$$

Considering the range of α_s, since $T_{s_0}\leq(d+4)n$,

$$\begin{aligned}
\alpha&=\sqrt{\frac{n\tau}{24L}}\geq\sqrt{\frac{T_{s_0}\tau}{24(d+4)L}}=\frac{1}{2}\cdot\frac{\tau}{d+4}\cdot\frac{1}{\sqrt{6nL\tau}}\cdot\sqrt{T_{s_0}(d+4)n}\\
&=\frac{\tau\gamma}{2}\cdot\sqrt{T_{s_0}(d+4)n}\geq\frac{\tau\gamma T_{s_0}}{2}.
\end{aligned}$$

Also note that, for any $T>1$ and $0\leq\delta T\leq1$, $(1+T\delta)\leq(1+\delta)^T\leq(1+2T\delta)$. If we set $\delta=\frac{\tau\gamma}{2}$ and $T=T_{s_0}$ here,

$$\delta T=\frac{\tau\gamma T_{s_0}}{2}\leq\alpha<1.$$

Then, we have

$$\begin{aligned}
1-(1-\alpha-p)(1+\frac{\tau\gamma}{2}) &= (1+\frac{\tau\gamma}{2})(\alpha+p-\frac{\tau\gamma}{2})+\frac{\tau^2\gamma^2}{4}\\
&\geq (1+\frac{\tau\gamma}{2})(\frac{\tau\gamma T_{s_0}}{2}+p-\frac{\tau\gamma}{2})\\
&= p(1+\frac{\tau\gamma}{2})(1+2(T_{s_0}-1)\cdot\frac{\tau\gamma}{2})\\
&\geq p(1+\frac{\tau\gamma}{2})^{T_{s_0}} = p\Gamma_{T_{s_0}}.
\end{aligned}$$

Hence, we obtain $\sum_{t=1}^{T_{s_0}}\theta_t \geq \Gamma_{T_{s_0}}\cdot\big[1-\alpha-p+p\sum_{t=1}^{T_s}\Gamma_{t-1}\big]$. Moreover, using Eq. (51) and the fact that $f_\mu(\tilde{x}^s)-f_\mu^*\geq 0$, we have

$$\begin{aligned}
&\Gamma_{T_{s_0}}\cdot\left[\frac{\gamma}{\alpha}\big[1-\alpha-p+p\sum_{t=1}^{T_s}\Gamma_{t-1}\big]\mathbb{E}_{\mathcal{F}_{T_s}}\big[f_\mu(\tilde{x}^s)-f_\mu^*\big]+\frac{1}{2}\mathbb{E}_{\mathcal{F}_{T_s}}\|x^s-x_\mu^*\|^2\right]\\
&\leq \frac{\gamma}{\alpha}\big[1-\alpha-p+p\sum_{t=1}^{T_s}\Gamma_{t-1}\big]\big[f_\mu(\tilde{x}^{s-1})-f_\mu^*\big]+\frac{1}{2}\|x^{s-1}-x_\mu^*\|^2+\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\Gamma_{t-1}\cdot\mu^2L(d+4)^2+\frac{\gamma}{\tau}\sum_{t=1}^{T_s}\Gamma_{t-1}\cdot E.
\end{aligned}$$

Applying this inequality iteratively for $s>s_0$, we obtain

$$\begin{aligned}
&\frac{\gamma}{\alpha}\big[1-\alpha-p+p\sum_{t=1}^{T_s}\Gamma_{t-1}\big]\mathbb{E}\big[f_\mu(\tilde{x}^s)-f_\mu^*\big]+\frac{1}{2}\mathbb{E}\|x^s-x_\mu^*\|^2\\
&\leq\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0}\left[\frac{\gamma}{\alpha}\big[1-\alpha-p+p\sum_{t=1}^{T_s}\Gamma_{t-1}\big]\mathbb{E}\big[f_\mu(\tilde{x}^{s_0})-f_\mu^*\big]+\frac{1}{2}\|x^{s_0}-x_\mu^*\|^2\right]\\
&\quad+\sum_{j=1}^{s-s_0}\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{j}\left[\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\Gamma_{t-1}\cdot\mu^2L(d+4)^2+\frac{\gamma}{\tau}\sum_{t=1}^{T_s}\Gamma_{t-1}\cdot E\right].
\end{aligned}$$

Note that $\frac{\gamma}{\alpha}\big[1-\alpha-p+p\sum_{t=1}^{T_s}\Gamma_{t-1}\big]\geq\frac{\gamma p}{\alpha}\sum_{t=1}^{T_s}\Gamma_{t-1}\geq\frac{\gamma pT_s}{\alpha}=\frac{\gamma pT_{s_0}}{\alpha}$ and $p=\frac{1}{2}$, the inequality above implies

$$\begin{aligned}
&\mathbb{E}\big[f_\mu(\tilde{x}^s)-f_\mu^*\big]\\
&\leq\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0}\left[\mathbb{E}\big[f_\mu(\tilde{x}^{s_0})-f_\mu^*\big]+\frac{\alpha}{\gamma T_{s_0}}\mathbb{E}\big[\|x^{s_0}-x_\mu^*\|^2\big]\right]+\sum_{j=1}^{s-s_0}\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{j}\left[2\mu^2L(d+4)^2+\frac{2\alpha}{\tau}\cdot E\right]\\
&\leq\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0}\left[\mathbb{E}\big[f_\mu(\tilde{x}^{s_0})-f_\mu^*\big]+\frac{\alpha}{\gamma T_{s_0}}\mathbb{E}\big[\|x^{s_0}-x_\mu^*\|^2\big]\right]+\frac{1}{\Gamma_{T_{s_0}}-1}\left[2\mu^2L(d+4)^2+\frac{2\alpha}{\tau}\cdot E\right].
\end{aligned}$$

Next, as

$$\Gamma_{T_{s_0}}=\big(1+\frac{\tau\gamma}{2}\big)^{T_{s_0}}\geq 1+\frac{\tau\gamma T_{s_0}}{2}\geq 1+\frac{\tau\gamma(d+4)n}{4}=1+\frac{1}{4}\cdot\sqrt{\frac{n\tau}{6L}}$$

and $\frac{2\alpha}{\tau}=\sqrt{\frac{n}{6L\tau}}$, we have that, for $s>s_0$,

$$\begin{aligned}
&\mathbb{E}\big[f_\mu(\tilde{x}^s)-f_\mu^*\big]\\
&\leq\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0}\left[\mathbb{E}\big[f_\mu(\tilde{x}^{s_0})-f_\mu^*\big]+\frac{\alpha}{\gamma T_{s_0}}\mathbb{E}\big[\|x^{s_0}-x_\mu^*\|^2\big]\right]+4\sqrt{\frac{6L}{n\tau}}\left[2\mu^2L(d+4)^2+\sqrt{\frac{n}{6L\tau}}\cdot E\right]\\
&=\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0}\left[\mathbb{E}\big[f_\mu(\tilde{x}^{s_0})-f_\mu^*\big]+\frac{\alpha}{\gamma T_{s_0}}\mathbb{E}\big[\|x^{s_0}-x_\mu^*\|^2\big]\right]+8\sqrt{\frac{6L}{n\tau}}\mu^2L(d+4)^2+\frac{4E}{\tau}\\
&\leq\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0}2\left[\mathbb{E}\big[f_\mu(\tilde{x}^{s_0})-f_\mu^*\big]+\frac{\alpha}{\gamma T_{s_0}}\cdot\frac{1}{2}\mathbb{E}\big[\|x^{s_0}-x_\mu^*\|^2\big]\right]+8\sqrt{\frac{6L}{n\tau}}\mu^2L(d+4)^2+\frac{4E}{\tau}.
\end{aligned}$$

Note that, since $n < \frac{6L}{\tau}$, we have $\frac{\alpha}{\gamma} = 12(d+4)L\alpha^2 = \frac{(d+4)\tau n}{2} \leq 3(d+4)L$. Finally, for $s > s_0$,

$$
\begin{aligned}
&\mathbb{E}\big[f_\mu(\tilde{x}^s) - f_\mu^*\big] \\
\leq& \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0} 2\left[\mathbb{E}\big[f_\mu(\tilde{x}^{s_0}) - f_\mu^*\big] + \frac{3(d+4)L}{T_{s_0}} \cdot \frac{1}{2}\mathbb{E}\big[\|x^{s_0} - x_\mu^*\|^2\big]\right] + 8\sqrt{\frac{6L}{n\tau}}\mu^2 L(d+4)^2 + \frac{4E}{\tau} \\
\leq& \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0} 2\left[\frac{1}{2^{s_0+1}}(d+4)D_0 + 2\mu^2 L(d+4)^2 + \frac{E}{2\tau}\right] + 8\sqrt{\frac{6L}{n\tau}}\mu^2 L(d+4)^2 + \frac{4E}{\tau} \\
\leq& \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0} \frac{(d+4)D_0}{2^{s_0}} + \left(8\sqrt{\frac{6L}{n\tau}} + 4\right)\mu^2 L(d+4)^2 + \frac{5E}{\tau} \\
=& \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0} \frac{(d+4)D_0}{2T_{s_0}} + \left(8\sqrt{\frac{6L}{n\tau}} + 4\right)\mu^2 L(d+4)^2 + \frac{5E}{\tau} \\
\leq& \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0} \frac{D_0}{n} + \left(8\sqrt{\frac{6L}{n\tau}} + 4\right)\mu^2 L(d+4)^2 + \frac{5E}{\tau} \\
=& \left(1 + \frac{1}{2(d+4)} \cdot \sqrt{\frac{\tau}{6nL}}\right)^{-T_{s_0}(s-s_0)} \frac{D_0}{n} + \left(8\sqrt{\frac{6L}{n\tau}} + 4\right)\mu^2 L(d+4)^2 + \frac{5E}{\tau} \\
\leq& \left(1 + \frac{1}{2(d+4)} \cdot \sqrt{\frac{\tau}{6nL}}\right)^{-\frac{n(d+4)(s-s_0)}{2}} \frac{D_0}{n} + \left(8\sqrt{\frac{6L}{n\tau}} + 4\right)\mu^2 L(d+4)^2 + \frac{5E}{\tau} \\
\leq& \left(1 + \frac{1}{4} \cdot \sqrt{\frac{n\tau}{6L}}\right)^{-(s-s_0)} \frac{D_0}{n} + \left(8\sqrt{\frac{6L}{n\tau}} + 4\right)\mu^2 L(d+4)^2 + \frac{5E}{\tau}.
\end{aligned}
$$

The second inequality is based on Eq. (49) and the fourth and fifth inequalities rely on $T_{s_0} \geq \frac{(d+4)n}{2}$. The last inequality comes from $1 + T\delta \leq (1+\delta)^T$ when $\delta \geq 0$. □

Now, we can derive Theorem 4 based on Lemma 19, Lemma 20, Lemma 21.

Proof of Theorem 4. To summarize, we have obtained

$$
\mathbb{E}\big[f_\mu(\tilde{x}^s) - f_\mu^*\big] := \begin{cases}
\frac{1}{2^{s+1}}(d+4)D_0 + 2\mu^2 L(d+4)^2 + \frac{E}{2\tau}, & 1 \leq s \leq s_0 \\
\left(\frac{4}{5}\right)^{s-s_0} \frac{D_0}{n} + 12\mu^2 L(d+4)^2 + \frac{5E}{\tau}, & s > s_0 \text{ and } n \geq \frac{6L}{\tau} \\
\left(1 + \frac{1}{4}\sqrt{\frac{n\tau}{6L}}\right)^{-(s-s_0)} \frac{D_0}{n} & s > s_0 \text{ and } n < \frac{6L}{\tau} \\
\qquad + \left(8\sqrt{\frac{6L}{n\tau}} + 4\right)\mu^2 L(d+4)^2 + \frac{5E}{\tau}, &
\end{cases} \tag{52}
$$

from Lemma 19, Lemma 20, Lemma 21. Hence, the proof of Theorem 4 is completed. □

We conclude by deriving the final complexity result, stated in the main paper.

Proof of Corollary 5. Using the same technique as for the proof of Corollary 3, we can make the error terms depending on E or μ vanish. In addition to $\mu \leq \mathcal{O}\big(\sqrt{\frac{\epsilon}{Ld}}\big)$ comeing from functional approximation error (see proof of Corollary 3), we also need $\mu = \mathcal{O}\big(\frac{\epsilon^{1/2}}{L^{1/2}d}\big)$ for the first two cases ($1 \leq s \leq s_0$ or $s > s_0$ and $n \geq \frac{6L}{\tau}$), $\mu = \mathcal{O}\big(\frac{n^{1/4}\tau^{1/4}\epsilon^{1/2}}{L^{3/4}d}\big)$ for the third case ($s > s_0$ and $n < \frac{6L}{\tau}$) and $\mu = \mathcal{O}\big(\frac{\tau^{1/2}\epsilon^{1/2}}{Ld^{3/2}}\big)$, $\nu = \mathcal{O}\big(\frac{\tau^{1/2}\epsilon^{1/2}}{Ld^{1/2}}\big)$ to ensure ϵ-optimality, $\frac{\epsilon}{4}$ more specifically. Hence, we can proceed as in (Lan et al., 2019) , neglecting the errors coming from the DFO framework (note that a similar procedure is adopted also in (Nesterov & Spokoiny, 2011) and (Liu et al., 2018b;a)) . For the first case ($1 \leq s \leq s_0$) the total number of function queries is given in Corollary 3. Then, in the second case ($s > s_0$ and $n \geq \frac{6L}{\tau}$), the algorithm run at most $S := \mathcal{O}\big\{ \log\big(\frac{(d+4)D_0}{\epsilon}\big)\big\}$ epochs to ensure the first error with ϵ-optimality. Thus, the total number of function queries in this case is bounded by

$$dnS + \sum_{s=1}^{S} T_s \leq dn + S(d+4)n = \mathcal{O}\left\{ dn \log\left(\frac{dD_0}{\epsilon}\right)\right\}. \tag{53}$$

Finally, in the last case ($s > s_0$ and $n < \frac{6L}{\tau}$) to achieve ϵ-error for the first term, the algorithm need to run at most $S^{'} := s_0 + \sqrt{\frac{6L}{n\tau}} \log\big(\frac{D_0}{n\epsilon}\big)$ epochs. Therefore, the total number of function queries in this case is bounded by

$$\begin{aligned}\sum_{s=1}^{S^{'}}(dn + T_s) &= \sum_{s=1}^{s_0}(dn + T_s) + (dn + T_{s_0})(S^{'} - s_0)\\ &\leq 2dns_0 + \big(dn + (d+4)n\big)\sqrt{\frac{6L}{n\tau}}\log\left(\frac{D_0}{n\epsilon}\right)\\ &= \mathcal{O}\left\{ dn\log(dn) + d\sqrt{\frac{nL}{\tau}}\log\left(\frac{D_0}{n\epsilon}\right)\right\}.\end{aligned} \tag{54}$$

□

C. Proofs for Section 5: the coordinate-wise variant of Algorithm 1

When we replace the gradient estimator $g_\mu(x, u, i)$ in Algorithm 1 with Eq. (3), the dependency on the problem dimension d gets better (Lemma 14 compared to Lemma 1), and the analysis looks more like the original Varag analysis (Lan et al., 2019), with the addition of DFO errors. However, we should notice that Eq. (3) requires d times computation per iteration compared to Eq. (2). From another point of view, choosing the gradient estimation in derivative-free optimization is a trade off between computation time and numerical accuracy.

The first lemma follows directly from Lemma 5 in (Lan et al., 2019) (we simplify it to the case with $V(z, x) = \frac{1}{2}\|z - x\|^2$, $X = \mathbb{R}^d$ and $h(x) = 0$). Note that this is very similar to Lemma 15, but the Lemma below is with respect to f rather than f_μ. Indeed, for this appendix we define

$$\delta_t := G_t - \nabla f(\underline{x}_t).$$

Also we recall that, to make the notation compact, we define

$$x_{t-1}^{+} := \frac{1}{1+\tau\gamma_s}(x_{t-1} + \tau\gamma_s\underline{x}_t), \qquad l_f(z, x) := f(z) + \langle \nabla f(z), x - z\rangle.$$

Lemma 22. *Consider the coordinate-wise variant of Algorithm 1. Assume **(A1)**. For any $x \in \mathbb{R}^d$, we have*

$$\begin{aligned}&\gamma_s[l_f(\underline{x}_t, x_t) - l_f(\underline{x}_t, x)]\\ &\qquad \leq \frac{\tau\gamma_s}{2}\|\underline{x}_t - x\|^2 + \frac{1}{2}\|x_{t-1} - x\|^2 - \frac{1+\tau\gamma_s}{2}\|x_t - x\|^2 - \frac{1+\tau\gamma_s}{2}\|x_t - x_{t-1}^{+}\|^2 - \gamma_s\langle\delta_t, x_t - x\rangle.\end{aligned}$$

which can be rewritten as

$$\begin{aligned}&\gamma_s\langle\nabla f(\underline{x}_t), x_t - x\rangle\\ &\qquad \leq \frac{\tau\gamma_s}{2}\|\underline{x}_t - x\|^2 + \frac{1}{2}\|x_{t-1} - x\|^2 - \frac{1+\tau\gamma_s}{2}\|x_t - x\|^2 - \frac{1+\tau\gamma_s}{2}\|x_t - x_{t-1}^{+}\|^2 - \gamma_s\langle\delta_t, x_t - x\rangle.\end{aligned}$$

The next lemma is similar to Lemma 6 in (Lan et al., 2019), but with some additional error terms, due to DFO framework.

Lemma 23. *Consider the coordinate-wise variant of Algorithm 1. Assume **(A1)**. Assume that $\alpha_s \in [0,1]$, $p_s \in [0,1]$ and $\gamma_s > 0$ satisfy*

$$1 + \tau\gamma_s - L\alpha_s\gamma_s > 0, \tag{55}$$

$$p_s - \frac{4L\alpha_s\gamma_s}{1 + \tau\gamma_s - L\alpha_s\gamma_s} \geq 0. \tag{56}$$

Under the expectation of i_t, we have

$$\begin{aligned}
&\mathbb{E}_{i_t}\left[\frac{\gamma_s}{\alpha_s}\big[f(\bar{x}_t) - f(x)\big] + \frac{(1+\tau\gamma_s)}{2}\|x_t - x\|^2\right] \\
&\leq \frac{\gamma_s}{\alpha_s}(1 - \alpha_s - p_s)\big[f(\bar{x}_{t-1}) - f(x)\big] + \frac{\gamma_s p_s}{\alpha_s}\big[f(\tilde{x}) - f(x)\big] + \frac{1}{2}\|x_{t-1} - x\|^2 \\
&\quad + \frac{\gamma_s}{\alpha_s} \cdot \frac{6\alpha_s\gamma_s\nu^2L^2d}{1 + \tau\gamma_s - L\alpha_s\gamma_s} - \frac{\gamma_s}{\alpha_s} \cdot \alpha_s\langle g_\nu(\underline{x}_t) - \nabla f(\underline{x}_t), x^+_{t-1} - x\rangle.
\end{aligned} \tag{57}$$

for any $x \in \mathbb{R}^d$.

Proof of Lemma 23. By the L-smoothness of f,

$$\begin{aligned}
f(\bar{x}_t) &\leq f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), \bar{x}_t - \underline{x}_t\rangle + \frac{L}{2}\|\bar{x}_t - \underline{x}_t\|^2 \\
&= (1 - \alpha_s - p_s)\big[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), \bar{x}_{t-1} - \underline{x}_t\rangle\big] + \alpha_s\big[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), x_t - \underline{x}_t\rangle\big] \\
&\quad + p_s\big[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] + \frac{L\alpha_s^2}{2}\|x_t - x^+_{t-1}\|^2.
\end{aligned}$$

The equality above holds because of the update rule of $\bar{x}_t$ in Algorithm 1 and Eq. (24). Then, applying Lemma 22 for the inequality above, we have

$$\begin{aligned}
&f(\bar{x}_t) \\
&\leq (1 - \alpha_s - p_s)\big[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), \bar{x}_{t-1} - \underline{x}_t\rangle\big] + \alpha_s\big[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), x - \underline{x}_t\rangle\big] \\
&\quad + \alpha_s\big[\frac{\tau}{2}\|\underline{x_t} - x\|^2 + \frac{1}{2\gamma_s}\|x_{t-1} - x\|^2 - \frac{1+\tau\gamma_s}{2\gamma_s}\|x_t - x\|^2 - \frac{1+\tau\gamma_s}{2\gamma_s}\|x_t - x^+_{t-1}\|^2 - \langle \delta_t, x_t - x\rangle\big] \\
&\quad + p_s\big[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] + \frac{L\alpha_s^2}{2}\|x_t - x^+_{t-1}\|^2 \\
&\leq (1 - \alpha_s - p_s)\big[f(\bar{x}_{t-1}) - \frac{\tau}{2}\|\bar{x}_{t-1} - \underline{x}_t\|^2\big] + \alpha_s\big[f(x) - \frac{\tau}{2}\|x - \underline{x}_t\|^2\big] \\
&\quad + \alpha_s\big[\frac{\tau}{2}\|\underline{x_t} - x\|^2 + \frac{1}{2\gamma_s}\|x_{t-1} - x\|^2 - \frac{1+\tau\gamma_s}{2\gamma_s}\|x_t - x\|^2\big] \\
&\quad + p_s\big[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] - \frac{\alpha_s}{2\gamma_s}(1 + \tau\gamma_s - L\alpha_s\gamma_s)\|x_t - x^+_{t-1}\|^2 - \alpha_s\langle \delta_t, x_t - x\rangle \\
&= (1 - \alpha_s - p_s)\big[f(\bar{x}_{t-1}) - \frac{\tau}{2}\|\bar{x}_{t-1} - \underline{x}_t\|^2\big] + \alpha_s\big[f(x) + \frac{1}{2\gamma_s}\|x_{t-1} - x\|^2 - \frac{1+\tau\gamma_s}{2\gamma_s}\|x_t - x\|^2\big] \\
&\quad + p_s\big[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] - \frac{\alpha_s}{2\gamma_s}(1 + \tau\gamma_s - L\alpha_s\gamma_s)\|x_t - x^+_{t-1}\|^2 - \alpha_s\langle \delta_t, x_t - x^+_{t-1}\rangle - \alpha_s\langle \delta_t, x^+_{t-1} - x\rangle \\
&\leq (1 - \alpha_s - p_s)\big[f(\bar{x}_{t-1}) - \frac{\tau}{2}\|\bar{x}_{t-1} - \underline{x}_t\|^2\big] + \alpha_s\big[f(x) + \frac{1}{2\gamma_s}\|x_{t-1} - x\|^2 - \frac{1+\tau\gamma_s}{2\gamma_s}\|x_t - x\|^2\big] \\
&\quad + p_s\big[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] + \frac{\alpha_s\gamma_s}{2(1 + \tau\gamma_s - L\alpha_s\gamma_s)}\|\delta_t\|^2 - \alpha_s\langle \delta_t, x^+_{t-1} - x\rangle.
\end{aligned} \tag{58}$$

The second inequality holds thanks to (strong) convexity of f. The last inequality follows from $b\langle u, v\rangle - \frac{a}{2}\|v\|^2 \leq \frac{b^2}{2a}\|u\|^2, \forall a > 0$; where we set $a = \frac{\alpha_s}{\gamma_s}(1 + \tau\gamma_s - L\alpha_s\gamma_s)$ and $b = -\alpha_s$, requiring $1 + \tau\gamma_s - L\alpha_s\gamma_s > 0$.

Note that $\delta_t = G_t - \nabla f(\underline{x}_t)$ for the coordinate-wise variant. Taking the expectation w.r.t i_t, according to Lemma 14,

$$
\begin{aligned}
& p_s\big[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] + \frac{\alpha_s\gamma_s}{2(1+\tau\gamma_s - L\alpha_s\gamma_s)}\mathbb{E}_{i_t}\big[\|\delta_t\|^2\big] - \alpha_s\mathbb{E}_{i_t}\big[\langle \delta_t, x_{t-1}^+ - x\rangle\big] \\
\leq\, & p_s\big[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] + \frac{6\alpha_s\gamma_s \cdot \nu^2 L^2 d}{1+\tau\gamma_s - L\alpha_s\gamma_s} + \frac{4\alpha_s\gamma_s L}{1+\tau\gamma_s - L\alpha_s\gamma_s}\big[f(\tilde{x}) - f(\underline{x}_t) - \langle \nabla f(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] \\
& - \alpha_s\big[\langle g_\nu(\underline{x}_t) - \nabla f(\underline{x}_t), x_{t-1}^+ - x\rangle\big] \\
=\, & \big(p_s - \frac{4\alpha_s\gamma_s L}{1+\tau\gamma_s - L\alpha_s\gamma_s}\big)\big[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\big] + \frac{6\alpha_s\gamma_s \cdot \nu^2 L^2 d}{1+\tau\gamma_s - L\alpha_s\gamma_s} + \frac{4\alpha_s\gamma_s L}{1+\tau\gamma_s - L\alpha_s\gamma_s} f(\tilde{x}) \\
& - \alpha_s\langle g_\nu(\underline{x}_t) - \nabla f(\underline{x}_t), x_{t-1}^+ - x\rangle \\
\leq\, & \big(p_s - \frac{4\alpha_s\gamma_s L}{1+\tau\gamma_s - L\alpha_s\gamma_s}\big)\big[f(\tilde{x}) - \frac{\tau}{2}\|\tilde{x} - \underline{x}_t\|^2\big] + \frac{6\alpha_s\gamma_s \cdot \nu^2 L^2 d}{1+\tau\gamma_s - L\alpha_s\gamma_s} + \frac{4\alpha_s\gamma_s L}{1+\tau\gamma_s - L\alpha_s\gamma_s} f(\tilde{x}) \\
& - \alpha_s\langle g_\nu(\underline{x}_t) - \nabla f(\underline{x}_t), x_{t-1}^+ - x\rangle \\
=\, & p_s f(\tilde{x}) - \big(p_s - \frac{4\alpha_s\gamma_s L}{1+\tau\gamma_s - L\alpha_s\gamma_s}\big) \cdot \frac{\tau}{2}\|\tilde{x} - \underline{x}_t\|^2 + \frac{6\alpha_s\gamma_s \cdot \nu^2 L^2 d}{1+\tau\gamma_s - L\alpha_s\gamma_s} - \alpha_s\langle g_\nu(\underline{x}_t) - \nabla f(\underline{x}_t), x_{t-1}^+ - x\rangle,
\end{aligned}
\tag{59}
$$

where the last inequality holds if $p_s - \frac{4\alpha_s\gamma_s L}{1+\tau\gamma_s - L\alpha_s\gamma_s} \geq 0$. Combining Eq. (58) with Eq. (59), we obtain

$$
\begin{aligned}
& \mathbb{E}_{i_t}\big[f(\bar{x}_t) + \frac{\alpha_s(1+\tau\gamma_s)}{2\gamma_s}\|x_t - x\|^2\big] \\
\leq\, & (1-\alpha_s-p_s)f(\bar{x}_{t-1}) + p_s f(\tilde{x}) + \alpha_s f(x) + \frac{\alpha_s}{2\gamma_s}\|x_{t-1} - x\|^2 + \frac{6\alpha_s\gamma_s \cdot \nu^2 L^2 d}{1+\tau\gamma_s - L\alpha_s\gamma_s} \\
& - \alpha_s\langle g_\nu(\underline{x}_t) - \nabla f(\underline{x}_t), x_{t-1}^+ - x\rangle - \frac{(1-\alpha_s-p_s)\tau}{2}\|\bar{x}_{t-1} - \underline{x}_t\|^2 - \big(p_s - \frac{4\alpha_s\gamma_s(d+4)L}{1+\tau\gamma_s - L\alpha_s\gamma_s}\big) \cdot \frac{\tau}{2}\|\tilde{x} - \underline{x}_t\|^2 \\
\leq\, & (1-\alpha_s-p_s)f(\bar{x}_{t-1}) + p_s f(\tilde{x}) + \alpha_s f(x) + \frac{\alpha_s}{2\gamma_s}\|x_{t-1} - x\|^2 + \frac{6\alpha_s\gamma_s \cdot \nu^2 L^2 d}{1+\tau\gamma_s - L\alpha_s\gamma_s} \\
& - \alpha_s\langle g_\nu(\underline{x}_t) - \nabla f(\underline{x}_t), x_{t-1}^+ - x\rangle.
\end{aligned}
$$

Multiplying both sides by $\frac{\gamma_s}{\alpha_s}$ and then rearranging the inequality, we finish the proof of this lemma, i.e. Eq. (57). □

C.1. Proof of Theorem 6

To proceed, as in the Gaussian smoothing case, we need a technical assumption:

(A2$_\nu$) Let $x^* \in \mathrm{argmin}_{x\in\mathbb{R}^d} f(x)$. For any epoch s of Algorithm 1, consider the inner-loop sequences $\{\underline{x}_t\}$ and $\{\bar{x}_t\}$. There exist a *finite* constant $Z < \infty$, potentially dependent on L and d, such that, for ν small enough,

$$
\sup_{s\geq 0} \max_{x\in\{\bar{x}_t\}\cup\{\underline{x}_t\}} \mathbb{E}\left[\|x - x^*\|\right] \leq Z.
$$

Again, as mentioned in the context of **(A2$_\mu$)**, it is possible to show that this assumption holds under the requirement that f is coercive. As for Lemma 17, thanks to **(A2$_\nu$)**, we can get an epoch-wise inequality of the coordinate-wise approach.

Lemma 24. *Consider the coordinate-wise variant of Algorithm 1. Assume (A1), (A2$_\nu$). Set $\{\theta_t\}$ to*

$$
\theta_t = \begin{cases} \frac{\gamma_s}{\alpha_s}(\alpha_s + p_s) & 1 \leq t \leq T_s - 1 \\ \frac{\gamma_s}{\alpha_s} & t = T_s \end{cases}
\tag{60}
$$

and define

$$
\mathcal{L}_s := \frac{\gamma_s}{\alpha_s} + (T_s - 1)\frac{\gamma_s(\alpha_s + p_s)}{\alpha_s};
\tag{61}
$$

$$\mathcal{R}_s := \frac{\gamma_s}{\alpha_s}(1-\alpha_s) + (T_s - 1)\frac{\gamma_s p_s}{\alpha_s}. \tag{62}$$

Under the conditions in Eq. (55) *and Eq.* (56)*, we have:*

$$\begin{aligned}\mathcal{L}_s\mathbb{E}\big[f(\tilde{x}^s) - f(x^*)\big] \leq{}& \mathcal{R}_s \cdot \big[f(\tilde{x}^{s-1}) - f(x^*)\big] + \big(\frac{1}{2}\|x^{s-1} - x^*\|^2 - \frac{1}{2}\|x^s - x^*\|^2\big)\\ &+ T_s \cdot \frac{\gamma_s}{\alpha_s} \cdot \frac{6\alpha_s\gamma_s\nu^2L^2d}{1 - L\alpha_s\gamma_s} + T_s \cdot \frac{\gamma_s}{\alpha_s}(2-\alpha_s)L\sqrt{d}Z\nu,\end{aligned}$$

where $x^* := \arg\min_{x\in\mathbb{R}^d} f(x)$.

Proof of Lemma 24. If we set $x = x^*$, Lemma 23 can be written as

$$\begin{aligned}&\mathbb{E}_{i_t}\left[\frac{\gamma_s}{\alpha_s}\big[f(\bar{x}_t) - f(x^*)\big] + \frac{1}{2}\|x_t - x^*\|^2\right]\\ \leq{}& \frac{\gamma_s}{\alpha_s}(1-\alpha_s-p_s)\big[f(\bar{x}_{t-1}) - f(x^*)\big] + \frac{\gamma_s p_s}{\alpha_s}\big[f(\tilde{x}) - f(x^*)\big] + \frac{1}{2}\|x_{t-1} - x^*\|^2\\ &+ \frac{\gamma_s}{\alpha_s} \cdot \frac{6\alpha_s\gamma_s\nu^2L^2d}{1 - L\alpha_s\gamma_s} - \frac{\gamma_s}{\alpha_s}\cdot\alpha_s\langle g_\nu(\underline{x}_t) - \nabla f(\underline{x}_t), x_{t-1}^+ - x^*\rangle.\end{aligned}$$

Summing up these inequalities over $t = 1, \ldots, T_s$, using the definition of θ_t and $\bar{x}_0 = \tilde{x}$, we get

$$\begin{aligned}\sum_{t=1}^{T_s}\theta_t\mathbb{E}\big[f(\bar{x}_t) - f(x^*)\big] \leq{}& \left[\frac{\gamma_s}{\alpha_s}(1-\alpha_s) + (T_s-1)\frac{\gamma_s p_s}{\alpha_s}\right]\cdot\big[f(\tilde{x}) - f(x^*)\big] + \left(\frac{1}{2}\|x_0 - x^*\|^2 - \frac{1}{2}\|x_{T_s} - x^*\|^2\right)\\ &+ T_s \cdot \frac{\gamma_s}{\alpha_s}\cdot\frac{6\alpha_s\gamma_s\nu^2L^2d}{1 - L\alpha_s\gamma_s} - \frac{\gamma_s}{\alpha_s}\sum_{t=1}^{T_s}\alpha_s\mathbb{E}\big[\langle g_\nu(\underline{x}_t) - \nabla f(\underline{x}_t), x_{t-1}^+ - x^*\rangle\big].\end{aligned}$$

Noticing that $\tilde{x}^s = \sum_{t=1}^{T_s}\big(\theta_t\bar{x}_t\big)/\sum_{t=1}^{T_s}\theta_t$, $\tilde{x} = \tilde{x}^{s-1}$, $x_0 = x^{s-1}$, $x_{T_s} = x^s$ and thanks to the convexity of f, the inequality above implies

$$\begin{aligned}\sum_{t=1}^{T_s}\theta_t\mathbb{E}\big[f(\tilde{x}^s) - f(x^*)\big] \leq{}& \left[\frac{\gamma_s}{\alpha_s}(1-\alpha_s) + (T_s-1)\frac{\gamma_s p_s}{\alpha_s}\right]\cdot\big[f(\tilde{x}^{s-1}) - f(x^*)\big] + \left(\frac{1}{2}\|x^{s-1} - x^*\|^2 - \frac{1}{2}\|x^s - x^*\|^2\right)\\ &+ T_s \cdot \frac{\gamma_s}{\alpha_s}\cdot\frac{6\alpha_s\gamma_s\nu^2L^2d}{1 - L\alpha_s\gamma_s} - \frac{\gamma_s}{\alpha_s}\sum_{t=1}^{T_s}\alpha_s\mathbb{E}\big[\langle g_\nu(\underline{x}_t) - \nabla f(\underline{x}_t), x_{t-1}^+ - x^*\rangle\big],\end{aligned}$$

which is equivalent to

$$\begin{aligned}\mathcal{L}_s\mathbb{E}\big[f(\tilde{x}^s) - f(x^*)\big] \leq{}& \mathcal{R}_s\cdot\big[f(\tilde{x}^{s-1}) - f(x^*)\big] + \left(\frac{1}{2}\|x^{s-1} - x^*\|^2 - \frac{1}{2}\|x^s - x^*\|^2\right)\\ &+ T_s \cdot \frac{\gamma_s}{\alpha_s}\cdot\frac{6\alpha_s\gamma_s\nu^2L^2d}{1 - L\alpha_s\gamma_s} - \frac{\gamma_s}{\alpha_s}\sum_{t=1}^{T_s}\alpha_s\mathbb{E}\big[\langle g_\nu(\underline{x}_t) - \nabla f(\underline{x}_t), x_{t-1}^+ - x^*\rangle\big].\end{aligned} \tag{63}$$

Now, let us look at the additional term $\langle g_\nu(\underline{x}_t) - \nabla f(\underline{x}_t), x_{t-1}^+ - x^*\rangle$, *which can not eliminated by expectation compared to gradient-based Varag* (Lan et al., 2019). According to Eq. (24) and the update rule of $\bar{x}_t$ in Algorithm 1, we have

$$\begin{aligned}\alpha_s(x_{t-1}^+ - x^*) &= \alpha_s x_t + \underline{x}_t - \bar{x}_t - \alpha_s x^*\\ &= -(1-\alpha_s-p_s)\bar{x}_{t-1} - p_s\tilde{x} + \underline{x}_t - \alpha_s x^*\\ &= -(1-\alpha_s-p_s)(\bar{x}_{t-1} - x^*) - p_s(\tilde{x} - x^*) + (\underline{x}_t - x^*)\\ &\leq (1-\alpha_s-p_s)||\bar{x}_{t-1} - x^*|| + p_s||\tilde{x} - x^*|| + ||\underline{x}_t - x^*||.\end{aligned}$$

Thanks to assumption **(A2$_\nu$)**, we have

$$\alpha_s\langle g_\nu(\underline{x}_t) - \nabla f(\underline{x}_t), x_{t-1}^+ - x^*\rangle \leq \|g_\nu(\underline{x}_t) - \nabla f(\underline{x}_t)\|\cdot\|\alpha_s(x_{t-1}^+ - x^*)\|$$

$$\leq (2-\alpha_s)L\sqrt{d}Z\nu. \tag{64}$$

The last inequality comes from Lemma 13. Combining the previous inequality with Eq. (63), we have

$$\begin{aligned}\mathcal{L}_s\mathbb{E}\big[f(\tilde{x}^s)-f(x^*)\big] \leq \mathcal{R}_s\cdot\big[f(\tilde{x}^{s-1})-f(x^*)\big] + \big(\frac{1}{2}\|x^{s-1}-x^*\|^2 - \frac{1}{2}\|x^s-x^*\|^2\big)\\ + T_s\cdot\frac{\gamma_s}{\alpha_s}\cdot\frac{6\alpha_s\gamma_s\nu^2L^2d}{1-L\alpha_s\gamma_s} + T_s\cdot\frac{\gamma_s}{\alpha_s}(2-\alpha_s)L\sqrt{d}Z\nu.\end{aligned}$$

□

Remark. *Here $\mathbb{E}_{i_t}\big[\delta_t\big] = g_\nu(\underline{x}_t) - \nabla f(\underline{x}_t) \neq 0$. Notice that the error terms in Eq. (64), i.e. $\langle g_\nu(\underline{x}_t)-\nabla f(\underline{x}_t), x^+_{t-1}-x^*\rangle$, is different from its counterpart in Eq. (34), i.e. $\langle \tilde{g}-\nabla f_\mu(\tilde{x}), x_t - x^*\rangle$.*

Then, we can derive Theorem 6 for convex and smooth f_i, based on Lemma 24. For convenience of the reader, we re-write the theorem here.

Theorem 6. Consider the coordinate-wise variant of Algorithm 1. Assume **(A1)** and **(A2$_\nu$)**. Let us denote $s_0 := \lfloor \log n\rfloor + 1$. Suppose the weights $\{\theta_t\}$ are set as in Eq. (10) and parameters $\{T_s\}$, $\{\gamma_s\}$, $\{p_s\}$ are set as

$$T_s = \begin{cases} 2^{s-1}, & s\leq s_0\\ T_{s_0}, & s > s_0\end{cases},\ \gamma_s = \tfrac{1}{12L\alpha_s},\quad p_s = \tfrac{1}{2},\text{ with} \tag{65}$$

$$\alpha_s = \begin{cases}\frac{1}{2}, & s\leq s_0\\ \frac{2}{s-s_0+4}, & s > s_0\end{cases}. \tag{66}$$

Then, we have

$$\mathbb{E}\big[f(\tilde{x}^s)-f^*\big] \leq \begin{cases}\dfrac{D_0'}{2^{s+1}} + \varsigma_1 + \varsigma_2, & 1\leq s\leq s_0\\ \dfrac{16D_0'}{n(s-s_0+4)^2} + \delta_s\cdot(\varsigma_1+\varsigma_2), & s > s_0\end{cases}$$

where $\varsigma_1 = \nu^2Ld$, $\varsigma_2 = 4L\sqrt{d}Z\nu$, $\delta_s = \mathcal{O}(s-s_0)$ and D_0' is defined as

$$D_0' := 2[f(x^0)-f(x^*)] + 6L\|x^0-x^*\|^2, \tag{67}$$

where x^* is any finite minimizer of f.

Proof of Theorem 6. Assumption Eq. (55) and Eq. (56) are satisfied since

$$1+\tau\gamma_s - L\alpha_s\gamma_s = 1-\frac{1}{12} > 0, \tag{68}$$

$$p_s - \frac{4L\alpha_s\gamma_s}{1+\tau\gamma_s-L\alpha_s\gamma_s} = \frac{1}{2}-\frac{1}{3}\cdot\frac{1}{1-\frac{1}{12}} > 0. \tag{69}$$

We define

$$w_s := \mathcal{L}_s - \mathcal{R}_{s+1}. \tag{70}$$

As in (Lan et al., 2019), if $1\leq s < s_0$,

$$w_s = \mathcal{L}_s - \mathcal{R}_{s+1} = \frac{\gamma_s}{\alpha_s}\big[1+(T_s-1)(\alpha_s+p_s)-(1-\alpha_s)-(2T_s-1)p_s\big] = \frac{\gamma_s}{\alpha_s}\big[T_s(\alpha_s-\gamma_s)\big] = 0;$$

else, if $s \geq s_0$,

$$\begin{aligned} w_s = \mathcal{L}_s - \mathcal{R}_{s+1} &= \frac{\gamma_s}{\alpha_s} - \frac{\gamma_{s+1}}{\alpha_{s+1}}(1-\alpha_{s+1}) + (T_{s_0}-1)\left[\frac{\gamma_s(\alpha_s+p_s)}{\alpha_s} - \frac{\gamma_{s+1}p_{s+1}}{\alpha_{s+1}}\right] \\ &= \frac{1}{48L} + \frac{(T_{s_0}-1)\left[2(s-s_0+4)-1\right]}{96L} > 0. \end{aligned}$$

Hence, $w_s \geq 0$ for all s. Using Lemma 24 iteratively,

$$\begin{aligned} &\mathcal{L}_s \mathbb{E}\big[f(\tilde{x}^s) - f(x^*)\big] \\ &\leq \mathcal{R}_1 \cdot \mathbb{E}\big[f(\tilde{x}^0) - f(x^*)\big] + \mathbb{E}\left[\frac{1}{2}\|x^0 - x^*\|^2 - \frac{1}{2}\|x^s - x^*\|^2\right] + \sum_{j=1}^{s} T_j \cdot \frac{\gamma_j}{\alpha_j} \cdot \frac{6\alpha_j\gamma_j\nu^2L^2d}{1-L\alpha_j\gamma_j} \\ &\quad + \sum_{j=1}^{s} T_j \cdot \frac{\gamma_j}{\alpha_j}(2-\alpha_j)L\sqrt{d}Z\nu \\ &= \frac{1}{6L}\big[f(\tilde{x}^0) - f(x^*)\big] + \frac{1}{2}\|x^0 - x^*\|^2 + \sum_{j=1}^{s} T_j \cdot \frac{\gamma_j}{\alpha_j} \cdot \frac{6\alpha_j\gamma_j\nu^2L^2d}{1-L\alpha_j\gamma_j} + \sum_{j=1}^{s} T_j \cdot \frac{\gamma_j}{\alpha_j}(2-\alpha_j)L\sqrt{d}Z\nu \\ &= \frac{1}{12L}D_0' + \sum_{j=1}^{s} \frac{1}{2}T_j \cdot \frac{\gamma_j}{\alpha_j} \cdot \frac{\nu^2Ld}{1-L\alpha_j\gamma_j} + \sum_{j=1}^{s} T_j \cdot \frac{\gamma_j}{\alpha_j}(2-\alpha_j)L\sqrt{d}Z\nu \\ &\leq \frac{1}{12L}D_0' + \sum_{j=1}^{s} \frac{1}{2}T_j \cdot \frac{\gamma_j}{\alpha_j} \cdot \nu^2Ld + \sum_{j=1}^{s} 2T_j \cdot \frac{\gamma_j}{\alpha_j}L\sqrt{d}Z\nu. \end{aligned} \tag{71}$$

The last equality holds since $\alpha_j\gamma_j = \frac{1}{12L}$. We proceed in two cases:

Case I: If $s \leq s_0$, $\mathcal{L}_s = \frac{2^{s+1}}{12L}$, $\mathcal{R}_s = \frac{2^s}{12L} = \frac{\mathcal{L}_s}{2}$, $\frac{\gamma_s}{\alpha_s} = \frac{1}{3L}$, $T_s = 2^{s-1}$. Hence, we have

$$\mathbb{E}\big[f(\tilde{x}^s) - f(x^*)\big] \leq \frac{1}{2^{s+1}}D_0' + \nu^2Ld + 4L\sqrt{d}Z\nu, \qquad 1 \leq s \leq s_0. \tag{72}$$

Case II: If $s > s_0$, we have

$$\begin{aligned} \mathcal{L}_s &= \frac{1}{12\alpha_s^2}\big[(T_s-1)\alpha_s + \frac{1}{2}(T_s+1)\big] \\ &= \frac{(s-s_0+4)^2}{48L} \cdot \big[(T_{s_0}-1)\alpha_s + \frac{1}{2}(T_{s_0}+1)\big] \\ &\geq \frac{(s-s_0+4)^2}{96L} \cdot (T_{s_0}+1) \\ &\geq \frac{n \cdot (s-s_0+4)^2}{192L}. \end{aligned}$$

where the last inequality holds since $T_{s_0} = 2^{\lfloor \log_2 n \rfloor} \geq \frac{n}{2}$, i.e. $2^{s_0} \geq n$. Hence, Eq. (71) implies

$$\mathbb{E}\big[f(\tilde{x}^s) - f(x^*)\big] \leq \frac{16D_0'}{n(s-s_0+4)^2} + \mathcal{O}(s-s_0) \cdot \nu^2Ld + \mathcal{O}(s-s_0) \cdot L\sqrt{d}Z\nu. \tag{73}$$

□

We can now derive the final complexity result.

Proof of Corollary 7. Using the same technique as for the proof of Corollary 3, we can make the error terms depending on ν vanishing. This requires $\nu = \mathcal{O}\big(\frac{\epsilon^{1/2}}{L^{1/2}d^{1/2}}\big)$, $\nu = \mathcal{O}\big(\frac{\epsilon}{Ld^{1/2}Z}\big)$ for the first case ($1 \leq s \leq s_0$) while $\nu = \mathcal{O}\big(\frac{n^{1/4}\epsilon^{3/4}}{L^{1/2}d^{1/2}D_0'^{1/4}}\big)$, $\nu = \mathcal{O}\big(\frac{n^{1/2}\epsilon^{3/2}}{Ld^{1/2}ZD_0'^{1/2}}\big)$ for the second case ($s > s_0$) to ensure ϵ-optimality, $\frac{\epsilon}{2}$ more specifically. Hence, we can proceed as in (Lan et al., 2019) , neglecting the errors coming from the DFO framework (note that a similar procedure is adopted also in (Nesterov & Spokoiny, 2011) and (Liu et al., 2018b;a)). If $n \geq \frac{D_0'}{\epsilon}$, we require

$$\frac{D_0'}{2^{s_0+1}} \leq \frac{D_0'}{2n} \leq \frac{\epsilon}{2}.$$

Therefore, the number of epochs can be bounded by $S_l = \min\left\{\log\left(\frac{D_0'}{\epsilon}\right), s_0\right\}$, achieving ϵ optimality inside Case I (see proof of Theorem 6). The total number of function queries is bounded by

$$d\left(nS_l + \sum_{s=1}^{S_l} T_s\right) = d \cdot \mathcal{O}\left\{\min\left(n\log\left(\frac{D_0'}{\epsilon}\right), n\log(n), n\right)\right\} = d \cdot \mathcal{O}\left\{\min\left(n\log\left(\frac{D_0'}{\epsilon}\right), n\right)\right\},$$

where the coefficient d corresponds to the number of function queries for each gradient estimation. All in all, the number of function queries is $\mathcal{O}\left\{dn\log\left(\frac{D_0'}{\epsilon}\right)\right\}$.

If $n < \frac{D_0'}{\epsilon}$ (Case II), we have $S_h = \left\lceil\sqrt{\frac{32D_0'}{n\epsilon}} + s_0 - 4\right\rceil$, ensuring the first term in Eq. (73) is not bigger than $\frac{\epsilon}{2}$. We can achieve ϵ optimality. Hence, the total number of function queries is

$$d\left[ns_0 + \sum_{s=1}^{s_0} T_s + (T_{s_0}+n)(S_h - s_0)\right] \leq d\left[\sum_{s=1}^{s_0} T_s + (T_{s_0}+n)S_h\right] = \mathcal{O}\left\{d\sqrt{\frac{nD_0'}{\epsilon}} + dn\log(n)\right\}.$$

□

C.2. Proof of Theorem 8

In this section, we consider f to be strongly convex, which we denoted as **(A3)**. We rewrite below Theorem 8, for convenience of the reader:

Theorem 8. Consider the coordinate-wise variant of Algorithm 1. Assume **(A1)**, **(A2$_\nu$)** and **(A3)**. Let us denote $s_0 := \lfloor\log n\rfloor + 1$ and assume that the weights $\{\theta_t\}$ are set to Eq. (30) if $1 \leq s \leq s_0$. Otherwise, they are set to

$$\theta_t = \begin{cases} \Gamma_{t-1} - (1-\alpha_s-p_s)\Gamma_t, & 1 \leq t \leq T_s - 1, \\ \Gamma_{t-1}, & t = T_s, \end{cases} \tag{74}$$

where $\Gamma_t = \left(1+\tau\gamma_s\right)^t$. If the parameters $\{T_s\}$, $\{\gamma_s\}$ and $\{p_s\}$ set to Eq. (14) with

$$\alpha_s = \begin{cases} \frac{1}{2}, & s \leq s_0, \\ \min\{\sqrt{\frac{n\tau}{12L}}, \frac{1}{2}\}, & s > s_0. \end{cases} \tag{75}$$

We obtain

$$\mathbb{E}\left[f(\tilde{x}^s) - f^*\right] \leq \begin{cases} \frac{1}{2^{s+1}}D_0' + \varsigma_1 + \varsigma_2, & 1 \leq s \leq s_0 \\ (4/5)^{s-s_0}\frac{D_0'}{n} + \varsigma_1 + \varsigma_2, & s > s_0 \text{ and } n \geq \frac{3L}{\tau} \\ \left(1+\frac{1}{4}\sqrt{\frac{n\tau}{3L}}\right)^{-(s-s_0)}\frac{D_0'}{n} + \left(2\sqrt{\frac{3L}{n\tau}}+1\right)(\varsigma_1+\varsigma_2), & s > s_0 \text{ and } n < \frac{3L}{\tau} \end{cases}$$

where $\varsigma_1 = 9\nu^2 Ld$, $\varsigma_2 = 24L\sqrt{d}Z\nu$ and D_0' is defined as in Eq. (16).

We start with a lemma.

Lemma 25. *Consider the coordinate-wise variant of Algorithm 1. Assume* **(A1), (A2$_\nu$)** *and* **(A3)**. *Under the choice of parameters from Theorem 8, we have*

$$\mathbb{E}_{i_t}\left[\frac{\gamma_s}{\alpha_s}\big[f(\bar{x}_t)-f^*\big]+\frac{(1+\tau\gamma_s)}{2}\|x_t-x^*\|^2\right]$$
$$\le \frac{\gamma_s}{\alpha_s}(1-\alpha_s-p_s)\big[f(\bar{x}_{t-1})-f^*\big]+\frac{\gamma_s p_s}{\alpha_s}\big[f(\tilde{x})-f^*\big]+\frac{1}{2}\|x_{t-1}-x^*\|^2+\frac{\gamma_s}{\alpha_s}\cdot\frac{3}{4}\nu^2Ld+\frac{\gamma_s}{\alpha_s}(2-\alpha_s)L\sqrt{d}Z\nu. \tag{76}$$

Proof of Lemma 25. For strongly convex f, Eq. (57) becomes,

$$\begin{aligned}&\mathbb{E}_{i_t}\left[\frac{\gamma_s}{\alpha_s}\big[f(\bar{x}_t)-f^*\big]+\frac{(1+\tau\gamma_s)}{2}\|x_t-x^*\|^2\right]\\ &\le \frac{\gamma_s}{\alpha_s}(1-\alpha_s-p_s)\big[f(\bar{x}_{t-1})-f^*\big]+\frac{\gamma_s p_s}{\alpha_s}\big[f(\tilde{x})-f^*\big]+\frac{1}{2}\|x_{t-1}-x^*\|^2\\ &\quad+\frac{\gamma_s}{\alpha_s}\cdot\frac{6\alpha_s\gamma_s\nu^2L^2d}{1+\tau\gamma_s-L\alpha_s\gamma_s}-\frac{\gamma_s}{\alpha_s}\cdot\alpha_s\langle g_\nu(\underline{x}_t)-\nabla f(\underline{x}_t),x_{t-1}^+-x\rangle\\ &\le \frac{\gamma_s}{\alpha_s}(1-\alpha_s-p_s)\big[f(\bar{x}_{t-1})-f^*\big]+\frac{\gamma_s p_s}{\alpha_s}\big[f(\tilde{x})-f^*\big]+\frac{1}{2}\|x_{t-1}-x^*\|^2\\ &\quad+\frac{\gamma_s}{\alpha_s}\cdot\frac{3}{4}\nu^2Ld+\frac{\gamma_s}{\alpha_s}\cdot(2-\alpha_s)L\sqrt{d}Z\nu.\end{aligned}$$

The last inequality holds when α_s and γ_s are as defined in Theorem 8 and Eq. (64). □

We divide the proof of Theorem 8 into three cases, corresponding to Lemma 26, Lemma 27, Lemma 28.

Lemma 26. *Consider the coordinate-wise variant of Algorithm 1. Assume* **(A1), (A2$_\nu$)** *and* **(A3)**. *Under the choice of parameters from Theorem 8, if $s\le s_0$, for any $x\in\mathbb{R}^d$ we have*

$$\mathbb{E}\big[f(\tilde{x}^s)-f^*\big]\le\frac{1}{2^{s+1}}D_0'+\frac{3}{2}\nu^2Ld+4L\sqrt{d}Z\nu,$$

where D_0' is defined in Eq. (16).

Proof of Lemma 26. For this case, $\alpha_s=p_s=\frac{1}{2}$, $\gamma_s=\frac{1}{6L}$, $T_s=2^{s-1}$. For Lemma 25, sum it up from $t=1$ to T_s, we have

$$\begin{aligned}&\sum_{t=1}^{T_s}\frac{\gamma_s}{\alpha_s}\mathbb{E}\big[f(\bar{x}_t)-f^*\big]+\frac{1}{2}\mathbb{E}\big[\|x_{T_s}-x^*\|^2\big]\\ &\le\frac{\gamma_s}{2\alpha_s}\cdot T_s\big[f(\tilde{x})-f^*\big]+\frac{1}{2}\|x_0-x^*\|^2+T_s\cdot\frac{\gamma_s}{\alpha_s}\cdot\frac{3}{4}\nu^2Ld+T_s\cdot\frac{\gamma_s}{\alpha_s}\cdot(2-\alpha_s)L\sqrt{d}Z\nu.\end{aligned}$$

Since f is convex, we have

$$\begin{aligned}&\frac{1}{3L}\cdot T_s\cdot\mathbb{E}\big[f(\tilde{x}^s)-f^*\big]+\frac{1}{2}\mathbb{E}\big[\|x^s-x^*\|^2\big]\\ &\le\frac{1}{6L}\cdot T_s\big[f(\tilde{x}^{s-1})-f^*\big]+\frac{1}{2}\|x^{s-1}-x^*\|^2+T_s\cdot\frac{1}{3L}\cdot\frac{3}{4}\nu^2Ld+T_s\cdot\frac{1}{3L}\cdot(2-\alpha_s)L\sqrt{d}Z\nu\\ &=\frac{1}{3L}\cdot T_{s-1}\big[f(\tilde{x}^{s-1})-f^*\big]+\frac{1}{2}\|x^{s-1}-x^*\|^2+T_s\cdot\frac{1}{3L}\cdot\frac{3}{4}\nu^2Ld+T_s\cdot\frac{1}{3L}\cdot(2-\alpha_s)L\sqrt{d}Z\nu,\end{aligned}$$

where $x_{T_s}=x^s$, $x_0=x^{s-1}$, $\tilde{x}=\tilde{x}^{s-1}$. From using the last inequality iteratively, we obtain

$$\frac{1}{3L}\cdot T_s\cdot\mathbb{E}\big[f(\tilde{x}^s)-f^*\big]+\frac{1}{2}\mathbb{E}\big[\|x^s-x^*\|^2\big]$$

$$\leq \frac{1}{3L}\cdot T_0\big[f(\tilde{x}^0)-f^*\big]+\frac{1}{2}\|x^0-x^*\|^2+\frac{1}{3L}\sum_{j=1}^{s}T_j\cdot\frac{3}{4}\nu^2Ld+\frac{1}{3L}\sum_{j=1}^{s}T_j\cdot(2-\alpha_j)L\sqrt{d}Z\nu$$

where $T_0=\frac{1}{2}$ is in accordance with the definition of $T_s=2^{s-1},\ s>0$. Hence, we obtain

$$\begin{aligned}&\mathbb{E}\big[f(\tilde{x}^s)-f^*\big]+\frac{3L}{T_s}\cdot\frac{1}{2}\mathbb{E}\big[\|x^s-x^*\|^2\big]\\ &\leq\frac{1}{2^s}\big[f(\tilde{x}^0)-f^*+3L\|x^0-x^*\|^2\big]+\frac{1}{T_s}\sum_{j=1}^{s}T_j\cdot\frac{3}{4}\nu^2Ld+\frac{1}{T_s}\sum_{j=1}^{s}T_j\cdot(2-\alpha_j)L\sqrt{d}Z\nu\\ &\leq\frac{1}{2^{s+1}}D_0'+\frac{3}{2}\nu^2Ld+4L\sqrt{d}Z\nu. \qquad (77)\end{aligned}$$

We conclude the proof by observing that $\frac{1}{2^{s-1}}\sum_{j=1}^{s}T_j\leq 2$ when $s\leq s_0$. □

Lemma 27. *Consider the coordinate-wise variant of Algorithm 1. Assume **(A1)**, **(A2$_\nu$)** and **(A3)**. Under the choice of parameters from Theorem 8, if $s\geq s_0$ and $n\geq\frac{3L}{\tau}$, then for any $x\in\mathbb{R}^d$ we have:*

$$\mathbb{E}\big[f(\tilde{x}^s)-f^*\big]\leq\left(\frac{4}{5}\right)^{s-s_0}\frac{D_0'}{n}+9\nu^2Ld+24L\sqrt{d}Z\nu.$$

Proof of Lemma 27. For this case, $\alpha_s=\alpha=p_s=\frac{1}{2}$, $\gamma_s=\gamma=\frac{1}{6L}$, $T_s=2^{s_0-1}$ when $s\geq s_0$. Based on Lemma 25, we have

$$\mathbb{E}_{i_t}\left[\frac{\gamma}{\alpha}\big[f(\bar{x}_t)-f^*\big]+\big(1+\tau\gamma\big)\cdot\frac{1}{2}\|x_t-x^*\|^2\right]\leq\frac{\gamma}{2\alpha}\big[f(\tilde{x})-f^*\big]+\frac{1}{2}\|x_{t-1}-x^*\|^2+\frac{\gamma}{\alpha}\cdot\frac{3}{4}\nu^2Ld+\frac{\gamma}{\alpha}\cdot2L\sqrt{d}Z\nu.$$

Multiplying both sides by $\Gamma_{t-1}=(1+\tau\gamma)^{t-1}$, we obtain

$$\begin{aligned}&\mathbb{E}_{i_t}\left[\frac{\gamma}{\alpha}\Gamma_{t-1}\big[f(\bar{x}_t)-f^*\big]+\frac{\Gamma_t}{2}\|x_t-x^*\|^2\right]\\ &\leq\frac{\gamma}{2\alpha}\Gamma_{t-1}\big[f(\tilde{x})-f^*\big]+\frac{\Gamma_{t-1}}{2}\|x_{t-1}-x^*\|^2+\frac{\gamma}{\alpha}\Gamma_{t-1}\cdot\frac{3}{4}\nu^2Ld+\frac{\gamma}{\alpha}\Gamma_{t-1}\cdot2L\sqrt{d}Z\nu.\end{aligned}$$

Since $\theta_t=\Gamma_{t-1}$, as defined in Eq. (17), the last inequality can be rewritten as

$$\mathbb{E}_{i_t}\left[\frac{\gamma}{\alpha}\theta_t\big[f(\bar{x}_t)-f^*\big]+\frac{\Gamma_t}{2}\|x_t-x^*\|^2\right]\leq\frac{\gamma}{2\alpha}\theta_t\big[f(\tilde{x})-f^*\big]+\frac{\Gamma_{t-1}}{2}\|x_{t-1}-x^*\|^2+\frac{\gamma}{\alpha}\theta_t\cdot\frac{3}{4}\nu^2Ld+\frac{\gamma}{\alpha}\theta_t\cdot2L\sqrt{d}Z\nu.$$

Summing up the inequality above from $t=1$ to T_s, we obtain

$$\begin{aligned}&\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\theta_t\mathbb{E}\big[f(\bar{x}_t)-f^*\big]+\frac{\Gamma_{T_s}}{2}\mathbb{E}\|x_{T_s}-x^*\|^2\\ &\leq\frac{\gamma}{2\alpha}\sum_{t=1}^{T_s}\theta_t\mathbb{E}\big[f(\tilde{x})-f^*\big]+\frac{1}{2}\|x_0-x^*\|^2+\frac{\gamma}{\alpha}\cdot\frac{3}{4}\nu^2Ld\sum_{t=1}^{T_s}\theta_t+\frac{\gamma}{\alpha}\cdot2L\sqrt{d}Z\nu\sum_{t=1}^{T_s}\theta_t,\end{aligned}$$

and then

$$\begin{aligned}&\frac{5}{4}\left[\frac{\gamma}{2\alpha}\sum_{t=1}^{T_s}\theta_t\mathbb{E}\big[f(\bar{x}_t)-f^*\big]+\frac{1}{2}\mathbb{E}\|x_{T_s}-x^*\|^2\right]\\ &\leq\frac{\gamma}{2\alpha}\sum_{t=1}^{T_s}\theta_t\mathbb{E}\big[f(\tilde{x})-f^*\big]+\frac{1}{2}\|x_0-x^*\|^2+\frac{\gamma}{\alpha}\cdot\frac{3}{4}\nu^2Ld\sum_{t=1}^{T_s}\theta_t+\frac{\gamma}{\alpha}\cdot2L\sqrt{d}Z\nu\sum_{t=1}^{T_s}\theta_t, \qquad (78)\end{aligned}$$

which is based on the fact that, for $s \geq s_0$, $\frac{n}{2} \leq T_s = T_{s_0} \leq n$, we have

$$\Gamma_{T_s} = \big(1+\tau\gamma\big)^{T_s} = \big(1+\tau\gamma\big)^{T_{s_0}} \geq 1+\tau\gamma\cdot T_{s_0} \geq 1+\tau\gamma\cdot\frac{n}{2} = 1+\frac{\tau n}{12L} \geq \frac{5}{4},$$

where the last inequality is conditioned on $n \geq \frac{3L}{\tau}$. Since $\tilde{x}^s = \sum_{t=1}^{T_s}(\theta_t\bar{x}_t)/\sum_{t=1}^{T_s}\theta_t$, $\tilde{x} = \tilde{x}^{s-1}$, $x_0 = x^{s-1}$, $x_{T_s} = x^s$ in the epoch s and thanks to the convexity of f, Eq. (78) implies

$$\begin{aligned}
&\frac{5}{4}\left[\frac{\gamma}{2\alpha}\mathbb{E}\big[f(\tilde{x}^s)-f^*\big] + \frac{1}{2\sum_{t=1}^{T_s}\theta_t}\mathbb{E}\|x^s-x^*\|^2\right] \\
&\leq \frac{\gamma}{2\alpha}\mathbb{E}\big[f(\tilde{x}^{s-1})-f^*\big] + \frac{1}{2\sum_{t=1}^{T_s}\theta_t}\|x^{s-1}-x^*\|^2 + \frac{\gamma}{\alpha}\cdot\frac{3}{4}\nu^2Ld + \frac{\gamma}{\alpha}\cdot 2L\sqrt{d}Z\nu.
\end{aligned}$$

Multiplying both sides with $\frac{2\alpha}{\gamma}$ and applying this inequality recursively for $s \geq s_0$, we obtain

$$\begin{aligned}
&\mathbb{E}\big[f(\tilde{x}^s)-f^*\big] + \frac{2\alpha}{\gamma\sum_{t=1}^{T_s}\theta_t}\cdot\frac{1}{2}\mathbb{E}\big[\|x^s-x^*\|^2\big] \\
&\leq \left(\frac{4}{5}\right)^{s-s_0}\left[\mathbb{E}\big[f(\tilde{x}^{s_0})-f^*\big] + \frac{2\alpha}{\gamma\sum_{t=1}^{T_s}\theta_t}\cdot\frac{1}{2}\mathbb{E}\big[\|x^{s_0}-x^*\|^2\big]\right] + \sum_{j=s_0+1}^{s}\left(\frac{4}{5}\right)^{s+1-j}\left[\frac{3}{2}\nu^2Ld + 4L\sqrt{d}Z\nu\right] \\
&\leq \left(\frac{4}{5}\right)^{s-s_0}\left[\mathbb{E}\big[f(\tilde{x}^{s_0})-f^*\big] + \frac{2\alpha}{\gamma T_{s_0}}\cdot\frac{1}{2}\mathbb{E}\big[\|x^{s_0}-x^*\|^2\big]\right] + 6\nu^2Ld + 16L\sqrt{d}Z\nu.
\end{aligned}$$

where the last inequality holds since $\sum_{j=s_0+1}^{s}\left(\frac{4}{5}\right)^{s+1-j} \leq \frac{4}{5}\cdot\frac{1}{1-\frac{4}{5}} = 4$ and $\sum_{t=1}^{T_s}\theta_t \geq T_s = T_{s_0}$. Hence

$$\begin{aligned}
&\mathbb{E}\big[f(\tilde{x}^s)-f^*\big] + \frac{2\alpha}{\gamma\sum_{t=1}^{T_s}\theta_t}\cdot\frac{1}{2}\mathbb{E}\big[\|x^s-x^*\|^2\big] \\
&\leq \left(\frac{4}{5}\right)^{s-s_0}\left[\mathbb{E}\big[f(\tilde{x}^{s_0})-f^*\big] + \frac{6L}{T_{s_0}}\cdot\frac{1}{2}\mathbb{E}\big[\|x^{s_0}-x^*\|^2\big]\right] + 6\nu^2Ld + 16L\sqrt{d}Z\nu \\
&\leq \left(\frac{4}{5}\right)^{s-s_0}2\left[\mathbb{E}\big[f(\tilde{x}^{s_0})-f^*\big] + \frac{3L}{T_{s_0}}\cdot\frac{1}{2}\mathbb{E}\big[\|x^{s_0}-x^*\|^2\big]\right] + 6\nu^2Ld + 16L\sqrt{d}Z\nu \\
&\leq \left(\frac{4}{5}\right)^{s-s_0}2\cdot\big[\frac{1}{2^{s_0+1}}D_0' + \frac{3}{2}\nu^2Ld + 4L\sqrt{d}Z\nu\big] + 6\nu^2Ld + 16L\sqrt{d}Z\nu \\
&\leq \left(\frac{4}{5}\right)^{s-s_0}\cdot\frac{D_0'}{2^{s_0}} + 9\nu^2Ld + 24L\sqrt{d}Z\nu \\
&= \left(\frac{4}{5}\right)^{s-s_0}\frac{D_0'}{2T_{s_0}} + 9\nu^2Ld + 24L\sqrt{d}Z\nu \\
&\leq \left(\frac{4}{5}\right)^{s-s_0}\frac{D_0'}{n} + 9\nu^2Ld + 24L\sqrt{d}Z\nu,
\end{aligned}$$

where the third inequality comes from Eq. (77) and the last inequality from the fact that $T_{s_0} \geq \frac{n}{2}$. □

Lemma 28. *Consider the coordinate-wise variant of Algorithm 1. Assume **(A1)**, **(A2$_\nu$)** and **(A3)**. If $s \geq s_0$ and $n < \frac{3L}{\tau}$, then for any $x \in \mathbb{R}^d$,*

$$\mathbb{E}\big[f(\tilde{x}^s)-f^*\big] \leq \left(1+\frac{1}{4}\sqrt{\frac{n\tau}{3L}}\right)^{-(s-s_0)}\frac{D_0'}{n} + \left(2\sqrt{\frac{3L}{n\tau}}+1\right)\left[3\nu^2Ld + 8L\sqrt{d}Z\nu\right].$$

Proof of Lemma 28. For this case, $\alpha_s = \alpha = \sqrt{\frac{n\tau}{12L}}$, $p_s = p = \frac{1}{2}$, $\gamma_s = \gamma = \frac{1}{\sqrt{12nL\tau}}$, $T_s = T_{s_0} = 2^{s_0-1}$ when $s \geq s_0$. Based on Lemma 25, we have

$$\mathbb{E}_{i_t}\left[\frac{\gamma}{\alpha}\big[f(\bar{x}_t)-f^*\big] + \frac{(1+\tau\gamma)}{2}\|x_t-x^*\|^2\right] \leq \frac{\gamma}{\alpha}(1-\alpha-p)\big[f(\bar{x}_{t-1})-f^*\big] + \frac{\gamma}{2\alpha}\big[f(\tilde{x})-f^*\big] + \frac{1}{2}\|x_{t-1}-x^*\|^2$$

$$+\frac{\gamma}{\alpha}\cdot\frac{3}{4}\nu^2 Ld+\frac{\gamma}{\alpha}\cdot(2-\alpha)L\sqrt{d}Z\nu.$$

Multiplying both sides by $\Gamma_{t-1}=(1+\tau\gamma)^{t-1}$, we obtain

$$\begin{aligned}\mathbb{E}_{i_t}\left[\frac{\gamma}{\alpha}\Gamma_{t-1}\big[f(\bar{x}_t)-f^*\big]+\frac{\Gamma_t}{2}\|x_t-x^*\|^2\right]&\le\frac{\Gamma_{t-1}\gamma}{\alpha}(1-\alpha-p)\big[f(\bar{x}_{t-1})-f^*\big]+\frac{\Gamma_{t-1}\gamma p}{\alpha}\big[f(\tilde{x})-f^*\big]\\&\quad+\frac{\Gamma_{t-1}}{2}\|x_{t-1}-x^*\|^2+\frac{\gamma}{\alpha}\Gamma_{t-1}\cdot\frac{3}{4}\nu^2Ld+\frac{\gamma}{\alpha}\Gamma_{t-1}\cdot(2-\alpha)L\sqrt{d}Z\nu.\end{aligned}$$

Summing up the inequality above from $t=1$ to T_s, we obtain

$$\begin{aligned}&\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\theta_t\mathbb{E}\big[f(\bar{x}_t)-f^*\big]+\frac{\Gamma_{T_s}}{2}\mathbb{E}\|x_{T_s}-x^*\|^2\\&\le\frac{\gamma}{\alpha}\Big[1-\alpha-p+p\sum_{t=1}^{T_s}\Gamma_{t-1}\Big]\mathbb{E}\big[f(\tilde{x})-f^*\big]+\frac{1}{2}\|x_0-x^*\|^2+\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\Gamma_{t-1}\cdot\frac{3}{4}\nu^2Ld+\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\Gamma_{t-1}\cdot(2-\alpha)L\sqrt{d}Z\nu.\end{aligned}$$

Since $\tilde{x}^s=\sum_{t=1}^{T_s}(\theta_t\bar{x}_t)/\sum_{t=1}^{T_s}\theta_t$, $\tilde{x}=\tilde{x}^{s-1}$, $x_0=x^{s-1}$, $x_{T_s}=x^s$ in the epoch s and the convexity of f_ν, it implies, for $s>s_0$,

$$\begin{aligned}&\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\theta_t\mathbb{E}\big[f(\tilde{x}^s)-f^*\big]+\frac{\Gamma_{T_{s_0}}}{2}\mathbb{E}\|x^s-x^*\|^2\\&\le\frac{\gamma}{\alpha}\Big[1-\alpha-p+p\sum_{t=1}^{T_s}\Gamma_{t-1}\Big]\mathbb{E}\big[f(\tilde{x}^{s-1})-f^*\big]+\frac{1}{2}\|x^{s-1}-x^*\|^2+\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\Gamma_{t-1}\cdot\frac{3}{4}\nu^2Ld+\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\Gamma_{t-1}\cdot(2-\alpha)L\sqrt{d}Z\nu.\end{aligned}\tag{79}$$

Moreover, we have

$$\begin{aligned}\sum_{t=1}^{T_{s_0}}\theta_t&=\Gamma_{T_{s_0}-1}+\sum_{t=1}^{T_{s_0}-1}\big(\Gamma_{t-1}-(1-\alpha-p)\Gamma_t\big)\\&=\Gamma_{T_{s_0}}(1-\alpha-p)+\sum_{t=1}^{T_{s_0}}\big(\Gamma_{t-1}-(1-\alpha-p)\Gamma_t\big)\\&=\Gamma_{T_{s_0}}(1-\alpha-p)+\big[1-(1-\alpha-p)(1+\tau\gamma)\big]\sum_{t=1}^{T_{s_0}}\Gamma_{t-1}.\end{aligned}$$

Considering the range of α_s, since $T_{s_0}\le n$,

$$\begin{aligned}\alpha=\sqrt{\frac{n\tau}{12L}}\ge\sqrt{\frac{T_{s_0}\tau}{12L}}&=\tau\cdot\frac{1}{\sqrt{12nL\tau}}\cdot\sqrt{T_{s_0}n}\\&=\tau\gamma\cdot\sqrt{T_{s_0}n}\ge\tau\gamma T_{s_0}.\end{aligned}$$

Also note that, for any $T>1$ and $0\le\delta T\le1$, $(1+T\delta)\le(1+\delta)^T\le(1+2T\delta)$. If we set $\delta=\tau\gamma$ and $T=T_{s_0}$ here,

$$\delta T=\tau\gamma T_{s_0}\le\alpha<1.$$

Then, we have

$$\begin{aligned}1-(1-\alpha-p)(1+\tau\gamma)&=(1+\tau\gamma)(\alpha+p-\tau\gamma)+\tau^2\gamma^2\\&\ge(1+\tau\gamma)(\tau\gamma T_{s_0}+p-\tau\gamma)\end{aligned}$$

$$
\begin{aligned}
&= p(1+\tau\gamma)(1+2(T_{s_0}-1)\cdot\tau\gamma) \\
&\geq p(1+\tau\gamma)^{T_{s_0}} = p\Gamma_{T_{s_0}}.
\end{aligned}
$$

Hence, we obtain $\sum_{t=1}^{T_{s_0}} \theta_t \geq \Gamma_{T_{s_0}} \cdot \big[1-\alpha-p+p\sum_{t=1}^{T_s}\Gamma_{t-1}\big]$. Moreover, thanks to Eq. (79) and $f(\tilde{x}^s)-f^*\geq 0$, the last inequality implies that

$$
\begin{aligned}
&\Gamma_{T_{s_0}} \cdot \left[\frac{\gamma}{\alpha}\big[1-\alpha-p+p\sum_{t=1}^{T_s}\Gamma_{t-1}\big]\mathbb{E}\big[f(\tilde{x}^s)-f^*\big]+\frac{1}{2}\mathbb{E}\|x^s-x^*\|^2\right] \\
\leq& \frac{\gamma}{\alpha}\big[1-\alpha-p+p\sum_{t=1}^{T_s}\Gamma_{t-1}\big]\mathbb{E}\big[f(\tilde{x}^{s-1})-f^*\big]+\frac{1}{2}\|x^{s-1}-x^*\|^2+\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\Gamma_{t-1}\cdot\frac{3}{4}\nu^2Ld \\
&+\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\Gamma_{t-1}\cdot(2-\alpha)L\sqrt{d}Z\nu.
\end{aligned}
$$

Applying this inequality iteratively for $s>s_0$, we obtain

$$
\begin{aligned}
&\frac{\gamma}{\alpha}\big[1-\alpha-p+p\sum_{t=1}^{T_s}\Gamma_{t-1}\big]\mathbb{E}\big[f(\tilde{x}^s)-f^*\big]+\frac{1}{2}\mathbb{E}\|x^s-x^*\|^2 \\
\leq& \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0}\left[\frac{\gamma}{\alpha}\big[1-\alpha-p+p\sum_{t=1}^{T_s}\Gamma_{t-1}\big]\mathbb{E}\big[f(\tilde{x}^{s_0})-f^*\big]+\frac{1}{2}\|x^{s_0}-x^*\|^2\right] \\
&+\sum_{j=1}^{s-s_0}\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^j\left[\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\Gamma_{t-1}\cdot\frac{3}{4}\nu^2Ld+\frac{\gamma}{\alpha}\sum_{t=1}^{T_s}\Gamma_{t-1}\cdot(2-\alpha)L\sqrt{d}Z\nu\right].
\end{aligned}
$$

Note that, since

$$
\frac{\gamma}{\alpha}\big[1-\alpha-p+p\sum_{t=1}^{T_s}\Gamma_{t-1}\big]\geq\frac{\gamma p}{\alpha}\sum_{t=1}^{T_s}\Gamma_{t-1}\geq\frac{\gamma p T_s}{\alpha}=\frac{\gamma p T_{s_0}}{\alpha}
$$

and $p=\frac{1}{2}$, the inequality above implies

$$
\begin{aligned}
&\mathbb{E}\big[f(\tilde{x}^s)-f^*\big] \\
\leq& \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0}\left[\mathbb{E}\big[f(\tilde{x}^{s_0})-f^*\big]+\frac{\alpha}{\gamma T_{s_0}}\mathbb{E}\big[\|x^{s_0}-x^*\|^2\big]\right]+\sum_{j=1}^{s-s_0}\big(\frac{1}{\Gamma_{T_{s_0}}}\big)^j\left[\frac{3}{2}\nu^2Ld+(4-2\alpha)L\sqrt{d}Z\nu\right] \\
\leq& \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0}\left[\mathbb{E}\big[f(\tilde{x}^{s_0})-f^*\big]+\frac{\alpha}{\gamma T_{s_0}}\mathbb{E}\big[\|x^{s_0}-x^*\|^2\big]\right]+\frac{1}{\Gamma_{T_{s_0}}-1}\left[\frac{3}{2}\nu^2Ld+(4-2\alpha)L\sqrt{d}Z\nu\right].
\end{aligned}
$$

As $\Gamma_{T_{s_0}}=\big(1+\tau\gamma\big)^{T_{s_0}}\geq 1+\tau\gamma T_{s_0}\geq 1+\frac{\tau\gamma n}{2}=1+\frac{1}{2}\cdot\sqrt{\frac{n\tau}{12L}}$, it implies, for $s>s_0$,

$$
\mathbb{E}\big[f(\tilde{x}^s)-f^*\big]\leq\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0}\left[\mathbb{E}\big[f(\tilde{x}^{s_0})-f^*\big]+\frac{\alpha}{\gamma T_{s_0}}\mathbb{E}\big[\|x^{s_0}-x^*\|^2\big]\right]+4\sqrt{\frac{3L}{n\tau}}\left[\frac{3}{2}\nu^2Ld+(4-2\alpha)L\sqrt{d}Z\nu\right].
$$

Note that, since $n<\frac{3L}{\tau}$, we have $\frac{\alpha}{\gamma}=12L\alpha^2\leq 3L$. Hence, for $s>s_0$, we have

$$
\begin{aligned}
&\mathbb{E}\big[f(\tilde{x}^s)-f^*\big] \\
\leq& \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0}2\left[\mathbb{E}\big[f(\tilde{x}^{s_0})-f^*\big]+\frac{3L}{T_{s_0}}\cdot\frac{1}{2}\mathbb{E}\big[\|x^{s_0}-x^*\|^2\big]\right]+4\sqrt{\frac{3L}{n\tau}}\left[\frac{3}{2}\nu^2Ld+(4-2\alpha)L\sqrt{d}Z\nu\right] \\
\leq& \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0}2\left[\frac{1}{2^{s_0+1}}D_0'+\frac{3}{2}\nu^2Ld+4L\sqrt{d}Z\nu\right]+4\sqrt{\frac{3L}{n\tau}}\left[\frac{3}{2}\nu^2Ld+(4-2\alpha)L\sqrt{d}Z\nu\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0} \frac{D_0'}{2^{s_0}} + (2\sqrt{\frac{3L}{n\tau}}+1)\left[3\nu^2 Ld + 8L\sqrt{d}Z\nu\right] \\
&= \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0} \frac{D_0'}{2T_{s_0}} + (2\sqrt{\frac{3L}{n\tau}}+1)\left[3\nu^2 Ld + 8L\sqrt{d}Z\nu\right] \\
&\leq \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0} \frac{D_0'}{n} + (2\sqrt{\frac{3L}{n\tau}}+1)\left[3\nu^2 Ld + 8L\sqrt{d}Z\nu\right] \\
&= \left(1+\frac{1}{2}\cdot\sqrt{\frac{\tau}{3nL}}\right)^{-T_{s_0}(s-s_0)} \frac{D_0'}{n} + (2\sqrt{\frac{3L}{n\tau}}+1)\left[3\nu^2 Ld + 8L\sqrt{d}Z\nu\right] \\
&\leq \left(1+\frac{1}{2}\cdot\sqrt{\frac{\tau}{3nL}}\right)^{-\frac{n(s-s_0)}{2}} \frac{D_0'}{n} + (2\sqrt{\frac{3L}{n\tau}}+1)\left[3\nu^2 Ld + 8L\sqrt{d}Z\nu\right] \\
&\leq \left(1+\frac{1}{4}\cdot\sqrt{\frac{n\tau}{3L}}\right)^{-(s-s_0)} \frac{D_0'}{n} + (2\sqrt{\frac{3L}{n\tau}}+1)\left[3\nu^2 Ld + 8L\sqrt{d}Z\nu\right],
\end{aligned}
$$

where the second inequality is based on Eq. (77) and the fourth and fifth inequalities rely on $T_{s_0} \geq \frac{n}{2}$. The last inequality comes from $1 + T\delta \leq (1+\delta)^T$ when $\delta \geq 0$. □

Now, we can finish the proof of Theorem 8:

Proof of Theorem 8. To summarize, we have obtained

$$
\mathbb{E}\big[f(\tilde{x}^s) - f^*\big] := \begin{cases}
\frac{1}{2^{s+1}}D_0' + \frac{3}{2}\nu^2 Ld + 4L\sqrt{d}Z\nu, & 1 \leq s \leq s_0 \\
\left(\frac{4}{5}\right)^{s-s_0}\frac{D_0'}{n} + 9\nu^2 Ld + 24L\sqrt{d}Z\nu, & s > s_0 \text{ and } n \geq \frac{3L}{\tau} \\
\left(1+\frac{1}{4}\sqrt{\frac{n\tau}{3L}}\right)^{-(s-s_0)}\frac{D_0'}{n} & s > s_0 \text{ and } n < \frac{3L}{\tau} \\
\quad +(2\sqrt{\frac{3L}{n\tau}}+1)\left[3\nu^2 Ld + 8L\sqrt{d}Z\nu\right], &
\end{cases} \tag{80}
$$

from Lemma 26, Lemma 27, Lemma 28. □

We conclude once again by proving the complexity result.

Proof of Corollary 9. Using the same technique as for the proof of Corollary 5, we can make the error terms depending on ν vanishing. It requires $\nu = \mathcal{O}\big(\frac{\epsilon^{1/2}}{L^{1/2}d^{1/2}}\big)$, $\nu = \mathcal{O}\big(\frac{\epsilon}{Ld^{1/2}Z}\big)$ for the first two cases ($1 \leq s \leq s_0$ or $s > s_0$ and $n \geq \frac{3L}{\tau}$) while we need to take the extra conditional number $\frac{L}{\tau}$ into account for the last case ($s > s_0$ and $n < \frac{3L}{\tau}$) to ensure ϵ-optimality, $\frac{\epsilon}{2}$ more specifically. Hence, we can proceed as in (Lan et al., 2019) , neglecting the errors coming from the DFO framework (note that a similar procedure is adopted also in (Nesterov & Spokoiny, 2011) and (Liu et al., 2018b;a)). For the first case above ($1 \leq s \leq s_0$), the total number of function queries is given in Theorem 6. In the second case ($s > s_0$ and $n \geq \frac{3L}{\tau}$), the algorithm runs at most $S := \mathcal{O}\big\{\log\big(\frac{D_0'}{\epsilon}\big)\big\}$ epochs to ensure ϵ-optimality. Thus, the total number of function queries in this case is bounded by

$$
dnS + \sum_{s=1}^{S} d \cdot T_s \leq dnS + dnS = \mathcal{O}\left\{dn\log\left(\frac{D_0'}{\epsilon}\right)\right\}. \tag{81}
$$

Finally, to achieve ϵ-error for the last case ($s > s_0$ and $n < \frac{3L}{\tau}$), our algorithm needs to run at most $S^{'} := s_0 +$

$\sqrt{\frac{3L}{n\tau}}\log\left(\frac{D_0'}{n\epsilon}\right)$ epochs. Therefore, the total number of function queries is bounded by

$$\begin{aligned}\sum_{s=1}^{S'}(dn+dT_s) &= \sum_{s=1}^{s_0}(dn+dT_s)+(dn+dT_{s_0})(S'-s_0)\\ &\le 2dns_0+\big(dn+dn\big)\sqrt{\frac{3L}{n\tau}}\log\left(\frac{D_0'}{n\epsilon}\right)\\ &= \mathcal{O}\Big\{dn\log(n)+d\sqrt{\frac{nL}{\tau}}\log\left(\frac{D_0'}{n\epsilon}\right)\Big\}. \qquad (82)\end{aligned}$$

□

D. Experiments

D.1. Parameter settings for Fig. 1

Here, we compare our method (Algorithm 1) with ZO-SVRG-Coord-Rand (Ji et al., 2019), with the accelerated method in (Nesterov & Spokoiny, 2011) and with a zero-order version of Katyusha inspired from (Shang et al., 2017) — which is a simplified version of the original algoerithm presented in (Allen-Zhu, 2017). We define this method in Algorithm 2.

Algorithm 2 Simplified ZO-Katyusha

```
Require: $x^0 \in \mathbb{R}^d, \{T_s\}, \{\gamma_s\}, \{\alpha_s\}$.
for $s = 1, 2, \dots, S$ do
    $\tilde{x} = \tilde{x}^{s-1}$, $x_0^s = y_0^s = \tilde{x}^{s-1}$;
    Pivotal ZO gradient $\tilde{g} = g_\nu(\tilde{x})$ using the coordinate-wise approach by Eq. (3).
    for $t = 1, 2, \dots, T_s$ do
        Pick $i_t$ uniformly at random from $\{1, \dots, n\}$;
        $G_t = g_\mu(x_{t-1}^s, u_t, i_t) - g_\mu(\tilde{x}, u_t, i_t) + \tilde{g}$;
        $y_t^s = y_{t-1}^s - \gamma_s G_t$;
        $x_t^s = \tilde{x} + \alpha_s(y_t^s - \tilde{x})$;
    end for
    $\tilde{x}^s = \frac{1}{T_s}\sum_{t=1}^{T_s} x_t^s$;
end for
Output: $\tilde{x}^S$
```

We recall some notation from the main paper.

- n is the data-set size;
- d is the problem dimension;
- b is the mini-batch size used to compute stochastic ZO-gradients.
- ν is the coordinate-smoothing parameter (see Section 3);
- μ is the Gaussian-smoothing parameter (see Section 3);
- $\{\alpha_s\}, \{\gamma_s\}, \{T_s\}$ and $\{\theta_t\}$ are parameters defined for ZO-Varag (Algorithm 1), which also appear in ZO-Katyusha (Algorithm 2). $\alpha_k, \gamma_k, \theta_k$ also appear in the algorithm by Nesterov & Spokoiny (2011), but have different definitions (see Eq. 60 in their paper);
- $\{p_s\}$ is the Katyusha momentum parameter in Algorithm 1, which can be seen as a *Katyusha momentum* even though it is defined differently in the simplified framework of (Shang et al., 2017) (see definition in the original paper by Allen-Zhu (2017));
- η is the step size for ZO-SVRG (Ji et al., 2019).

Next, we specify some parameter settings used for the experiments in Fig. 1. What is not specified here directly appears in the corresponding figures.

- $\mu = \nu = 0.001$.
- T_s are set as in Theorem 2.
- In Algorithm 1 and Algorithm 2 we set α_s according to Eq. (9) and Eq. (13) in the main paper. For the accelerated method by (Nesterov & Spokoiny, 2011), we used the choice of α_k reccomended in their paper.
- Note that, for both Algorithm 1 and Algorithm 2, the gradient estimate G_t is actually multiplied [6] by $\alpha_s \gamma_s$. Hence, $\alpha_s \gamma_s$ acts like a step-size. Therefore, in ZO-SVRG, we choose the *equivalent stepsize* $\eta_s = \alpha_s \gamma_s$. Also note that, as one can note in Eq. (8), $\alpha_s \gamma_s$ is actually constant and inversely proportional to d (see also next bullet-point).
- We choose γ_s such that $\eta = \alpha_s \gamma_s = 0.001 \cdot b/d$ for logistic regression and $\eta = \alpha_s \gamma_s = b/d$ for ridge regression when testing on the diabetes dataset (python sklearn). For the ijcnn1 dataset (python LIBSVM), we instead choose γ_s such that $\eta = \alpha_s \gamma_s = 0.1 \cdot b/d$ for logistic regression and $\eta = \alpha_s \gamma_s = 0.001 \cdot b/d$ for ridge regression.
- We pick $p_s = 0.5$, as specified in Eq. (8).

D.2. Additional experiments

Next, we discuss potential variations of the parameters discussed in the last subsection.

Options for pivotal point. We tested two options for pivot computation in Algorithm 1:

$$\text{Option I: } \tilde{x} = \tilde{x}^{s-1} = \tilde{x}^s = \textstyle\sum_{t=1}^{T_s} (\theta_t \bar{x}_t) / \big(\sum_{t=1}^{T_s} \theta_t \big) \text{ (as used in our analysis),} \quad \text{or} \quad \text{Option II: } \tilde{x} = \bar{x}^{s-1} \,.$$

In addition to the experimental results on the ijcnn1 dataset (LIBSVM) provided in Fig. 2, we also provide results on the diabetes dataset (sklearn) here: from Fig. 4, we observe that Option II also achieves faster convergence than Option I on the diabetes dataset in practice. Overall, our empirical evidence seems to indicate that Option II works better than Option I.

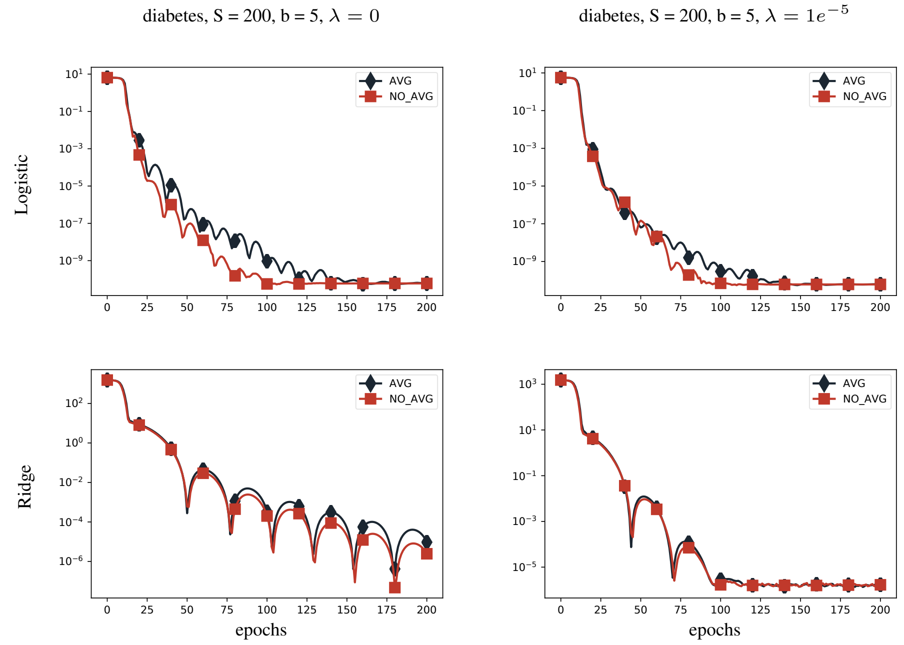

Figure 4. ZO-Varag, averaging (Option I) vs. no-averaging (Option II).

[6] In Algorithm 1, this is actually $\alpha_s \gamma_s / (1 + \mu \gamma_s) \approx \alpha_s \gamma_s$.

Effect of the momentum p_s. The effect of p_s (a.k.a Katyusha momentum) varies depending on the data set. From Fig. 5 and Fig. 6, we find that increasing values of p can either accelerate or slow down the convergence of the algorithm. Moreover, the algorithm may not converge when $p < 0.5$, since the constraint from Eq. (26) is not guaranteed anymore (recall the proof we provide is based on $p = 0.5$).

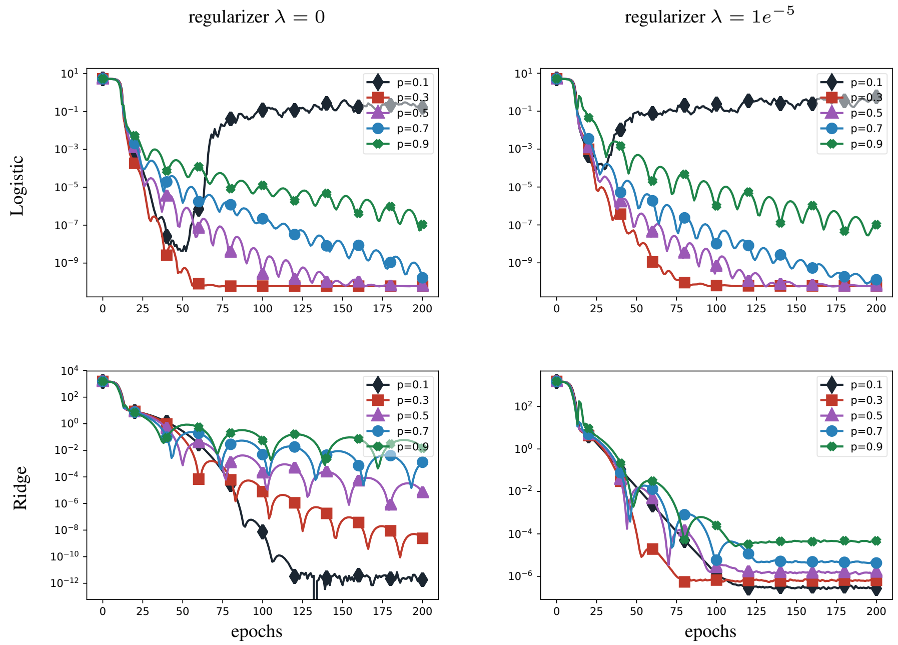

Figure 5. Effect of p on the diabetes dataset. Recall that our theoritical guarantees hold for $p = 0.5$.

Effect of the step-size $\alpha_s\gamma_s$. As discussed in Fig. 3 from the main paper, we find that the suboptimality stalling effect is related to the magnitude of the regularizer λ, which influences the strong-convexity constant of the objective function. Here, we show how such stalling effect of ZO-Varag can be controlled by tuning the step size $\alpha_s\gamma_s$. From Fig. 7, we see that the final suboptimality decreases if we decrease the magnitude of the step size $\alpha_s\gamma_s$, which however also affects speed of convergence.

Effect of the smoothing parameter ν. In this test, we set steps as the biggest step for each scenario in Fig. 7 as we only care about the stalling effects. In Fig. 8, we verify that the final error our ZO algorithm is dependent on the smoothing parameter ν at the pivotal point, i.e. smaller ν yields smaller error deviating from the optimum. However, we also find that this effect varies depending on the datasets and models being used, and is sometimes negligible: the logistic regression is sensitive to the values of the smoothing parameters, while the ridge regression is not. Note that, as expected, μ does not influence the steady-state error.

Comparison with the Coordinate-wise Variant Finally, we also provide a preliminary test between the ZO-Varag algorithm and its coordinate-wise variant which is introduced in Section 5. Although the length of inner loops are not the same for these two algorithms, see different definitions of s_0 in Theorem 2, 4, 6, 8, we only need to compare the function values at the pivotal points as the function queries are the same inside each inner loop after s_0 iterations (defined in Theorem 2). The experiments are carried out in Figure 9 and Figure 10, and show that there is almost no difference between the performance of ZO-Varag and the performance of its coordinate-wise variant, except the magnitude of stalling errors. This comes from the fact that the step size for the coordinate-wise variant is d times larger than that for ZO-Varag.

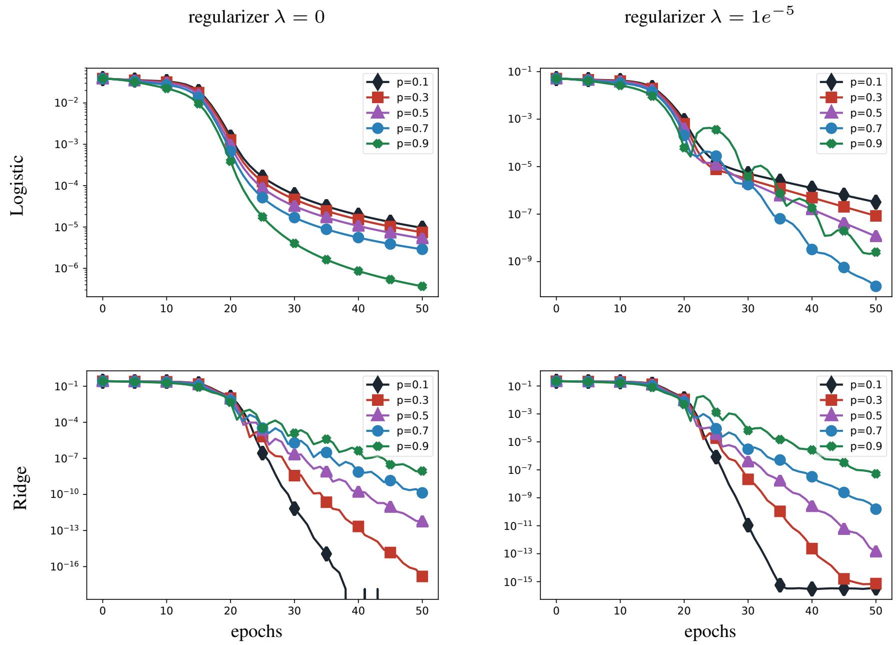

Figure 6. Effect of p on the ijcnn1 dataset. Recall that our theoritical guarantees hold for $p = 0.5$.

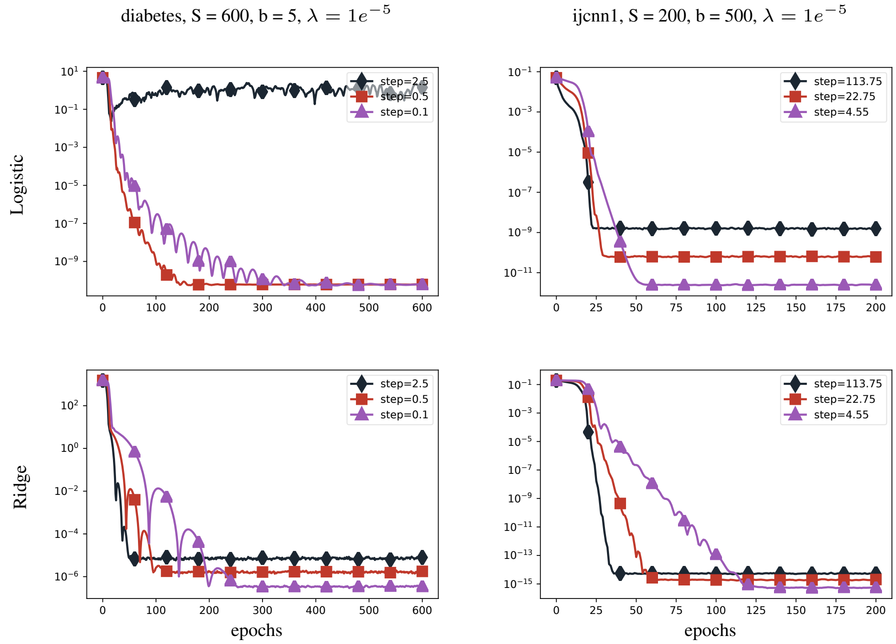

Figure 7. ZO-Varag with varying step-sizes= $\alpha_s \gamma_s$.

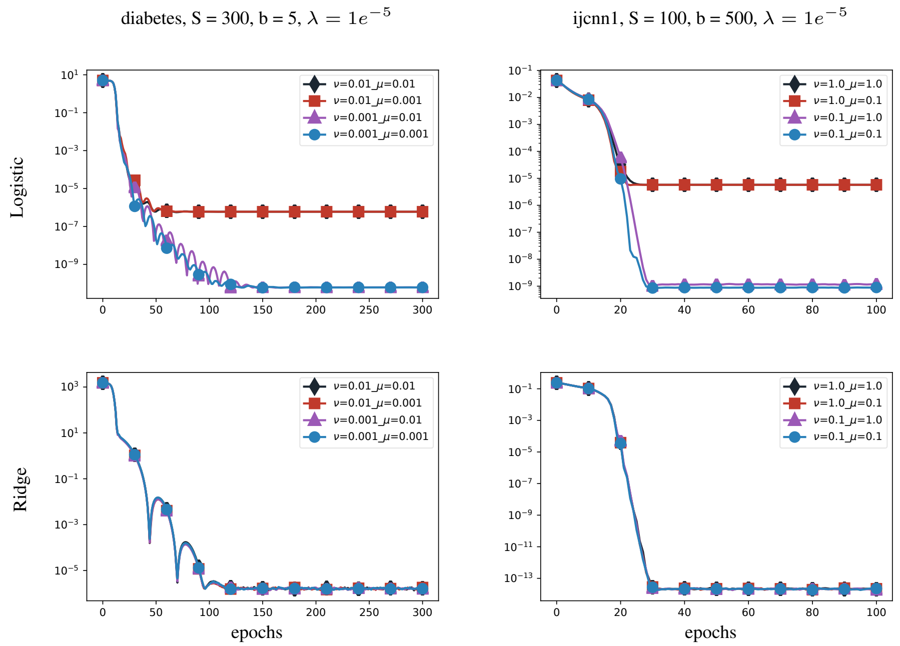

Figure 8. ZO-Varag, varying smoothing parameter μ and coordinate-wise paramater ν.

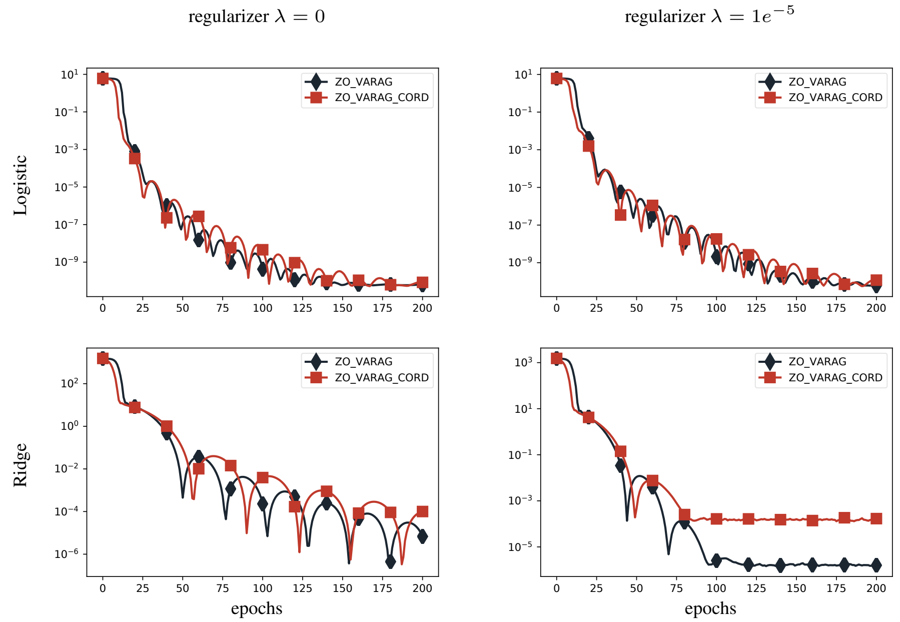

Figure 9. ZO-Varag vs. Coordinate-wise Variant of ZO-Varag (Diabetes)

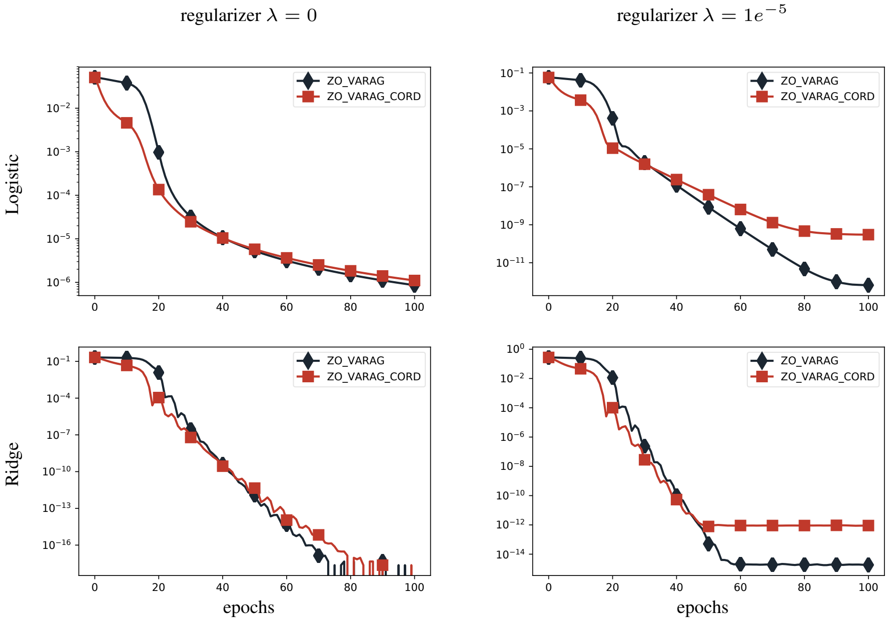

Figure 10. ZO-Varag vs. Coordinate-wise Variant of ZO-Varag (ijcnn1)