

## A SPECIFICATIONS OF EXPERIMENTS

The distributionally robust optimization problem is formulated as follows:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \sum_{i=1}^n y_i f_i(x) - r(y),$$

where  $\mathcal{X} = \{x \in \mathbb{R}^d\}$ ,  $\mathcal{Y} = \{y \in \mathbb{R}^n \mid \sum_{i=1}^n y_i = 1, y_i \geq 0, i = 1, \dots, n\}$ ,  $r(y) = 10 \sum_{i=1}^n (y_i - 1/n)^2$ ,  $f_i(x) = \phi(l(x))$  where  $\phi(\theta) = 2 \log(1 + \frac{\theta}{2})$ ,  $l(x; s, z) = \log(1 + \exp(-zx^\top s))$ , and  $(s, z)$  are the feature and label pair of a data sample. It can be seen that the problem is a min-max problem with  $d_1 = d$  and  $d_2 = n$ . Since the distributionally robust optimization aims at an unbalance dataset, we pick the samples from the original dataset and set the ratio between the number of negative labeled samples and the number of positive labeled samples to be 1 : 4. Since the maximization of  $y$  is a constrained optimization problem, we incorporate a projection step after updates of  $y$  for all algorithms.

The details of the datasets used for the comparison between SREDA and SREDA-Boost are listed in Table 2.

Table 2: Datasets used for the comparison between SREDA and SREDA-Boost

Datasets	# of samples	# of features	# Pos: # Neg
mushrooms	2000	112	1:4
w8a	5000	300	1:4
a9a	8000	123	1:4

The details of the datasets used for the comparison among zeroth-order algorithms are listed in Table 3.

Table 3: Datasets used for the comparison among zeroth-order algorithms

Datasets	# of samples	# of features	# Pos: # Neg
mushrooms	200	112	1:4
w8a	100	300	1:4
a9a	150	123	1:4

## B CONVERGENCE ANALYSIS OF SREDA-BOOST

### B.1 PRELIMINARIES

We first provide useful inequalities in convex optimization [Nesterov \(2013\)](#); [Polyak \(1963\)](#) and auxiliary lemmas from [Fang et al. \(2018\)](#); [Luo et al. \(2020\)](#).

**Lemma 1** ([Nesterov \(2013\)](#), [Polyak \(1963\)](#)). *Suppose  $h(\cdot)$  is convex and has  $\ell$ -Lipschitz gradient. Then, we have*

$$\langle \nabla h(w) - \nabla h(w'), w - w' \rangle \geq \frac{1}{\ell} \|\nabla h(w) - \nabla h(w')\|_2^2. \quad (3)$$

**Lemma 2** ([Nesterov \(2013\)](#), [Polyak \(1963\)](#)). *Suppose  $h(\cdot)$  is  $\mu$ -strongly convex and has  $\ell$ -Lipschitz gradient. Let  $w^*$  be the minimizer of  $h$ . Then for any  $w$  and  $w'$ , we have the following inequalities hold.*

$$\langle \nabla h(w) - \nabla h(w'), w - w' \rangle \geq \frac{\mu\ell}{\mu + \ell} \|w - w'\|_2^2 + \frac{1}{\mu + \ell} \|\nabla h(w) - \nabla h(w')\|_2^2, \quad (4)$$

$$\|\nabla h(w) - \nabla h(w')\|_2 \geq \mu \|w - w'\|_2, \quad (5)$$

$$2\mu(h(w) - h(w')) \leq \|\nabla h(w)\|_2^2. \quad (6)$$

The following structural lemma developed in Lin et al. (2019) provides further information about  $\Phi$  for nonconvex-strongly-concave min-max optimization.

**Lemma 3** (Lin et al. (2019), Lemma 4.3). *Under Assumption 2 and 3, the function  $\Phi(\cdot) = \max_{y \in \mathbb{R}^{d_2}} f(\cdot, y)$  is  $(\kappa + 1)\ell$ -gradient Lipschitz with  $\nabla \Phi(x) = \nabla_x f(x, y^*(x))$  where  $y^*(\cdot) = \arg\max_{y \in \mathbb{R}^{d_2}} f(\cdot, y)$ , and  $y^*(\cdot)$  is  $\kappa$ -Lipschitz.*

We let  $L \triangleq (1 + \kappa)\ell$  denote the Lipschitz constant of  $\nabla \Phi(x)$ .

**Lemma 4** (Fang et al. (2018), Lemma 2). *Suppose Assumption 4 hold. For any  $(x, y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$  and sample batch  $\{\xi_1, \dots, \xi_S\}$ , let  $v = \frac{1}{S} \sum_{i=1}^S \nabla_x F(x, y, \xi_i)$  and  $u = \frac{1}{S} \sum_{i=1}^S \nabla_y F(x, y, \xi_i)$ . We have*

$$\mathbb{E}[\|v - \nabla_x f(x, y)\|_2^2] + \mathbb{E}[\|u - \nabla_y f(x, y)\|_2^2] \leq \frac{\sigma^2}{S}.$$

**Lemma 5** (Fang et al. (2018), Lemma 1). *Let  $\mathcal{V}_t$  be an estimator of  $\mathcal{B}(z_t)$  as*

$$\mathcal{V}_t = \mathcal{B}_{S_*}(z_t) - \mathcal{B}_{S_*}(z_{t-1}) + \mathcal{V}_{t-1},$$

where  $\mathcal{B}_{S_*} = \frac{1}{|S_*|} \sum_{B_i \in S_*} B_i$  satisfies

$$\mathbb{E}[\mathcal{B}_i(z_t) - \mathcal{B}_i(z_{t-1}) | z_0, \dots, z_{t-1}] = \mathbb{E}[\mathcal{V}_t - \mathcal{V}_{t-1} | z_0, \dots, z_{t-1}].$$

For all  $k = 1, \dots, K$ , we have

$$\mathbb{E}[\|\mathcal{V}_t - \mathcal{V}_{t-1} - (\mathcal{B}_{S_*}(z_t) - \mathcal{B}_{S_*}(z_{t-1}))\|_2^2] \leq \frac{1}{S_*} \mathbb{E}[\|\mathcal{B}_i(z_t) - \mathcal{B}_i(z_{t-1})\|_2^2 | z_0, \dots, z_{t-1}],$$

and

$$\mathbb{E}[\|\mathcal{V}_t - \mathcal{B}(z_t) | z_0, \dots, z_{t-1}\|_2^2] \leq \|\mathcal{V}_{t-1} - \mathcal{B}(z_{t-1})\|_2^2 + \frac{1}{|S_*|} \mathbb{E}[\|\mathcal{B}_i(z_t) - \mathcal{B}_i(z_{t-1})\|_2^2 | z_0, \dots, z_{t-1}].$$

Furthermore, if  $\mathcal{B}_i$  is  $L$ -Lipschitz continuous in expectation, we have

$$\mathbb{E}[\|\mathcal{V}_t - \mathcal{B}(z_t) | z_0, \dots, z_{t-1}\|_2^2] \leq \|\mathcal{V}_{t-1} - \mathcal{B}(z_{t-1})\|_2^2 + \frac{L^2}{|S_*|} \mathbb{E}[\|z_t - z_{t-1}\|_2^2 | z_0, \dots, z_{t-1}].$$

## B.2 INITIALIZATION BY ISARAH

We present the detailed procedure of iSARAH in Algorithm 3, which is used to initialize  $y_0$  in SREDA-Boost (line 3 of Algorithm 1). We consider the following convex optimization problem:

$$\min_{w \in \mathbb{R}^d} p(w) \triangleq \mathbb{E}_\xi [P(w; \xi)], \quad (7)$$

where  $P$  is average  $\ell$ -gradient Lipschitz and convex,  $p$  is  $\mu$ -strongly convex, and  $\xi$  is a random vector.

---

### Algorithm 3 iSARAH

---

```

1: Input:  $\tilde{w}_0$ , learning rate  $\gamma > 0$ , inner loop size  $I$ , batch size  $B_1$  and  $B_2$ 
2: for  $t = 1, 2, \dots, T$  do
3:    $w_0 = \tilde{w}_{t-1}$ 
4:   draw  $B_1$  samples  $\{\xi_1, \dots, \xi_{B_1}\}$ 
5:    $v_0 = \frac{1}{B_1} \sum_{i=1}^{B_1} \nabla P(w_0, \xi_i)$ 
6:    $w_1 = w_0 + \gamma v_0$ 
7:   for  $k = 1, 2, \dots, I - 1$  do
8:     Draw minibatch sample  $\mathcal{M} = \{\xi_1, \dots, \xi_{B_2}\}$ 
9:      $v_k = v_{k-1} + \frac{1}{B_2} \sum_{i=1}^{B_2} \nabla P(w_k, \xi_i) - \frac{1}{B_2} \sum_{i=1}^{B_2} \nabla P(w_{k-1}, \xi_i)$ 
10:     $w_{k+1} = w_k - \gamma v_k$ 
11:   end for
12:    $\tilde{w}_t$  chosen uniformly at random from  $\{w_k\}_{k=0}^I$ 
13: end for

```

---

We have the following convergence result by using iSARAH to solve the problem in eq. (7).

**Lemma 6** (Nguyen et al. (2018), Corollary 4). *Consider Algorithm 3. Set  $\gamma = \Theta(\ell^{-1})$ ,  $B_1 = \Theta(\epsilon^{-1})$ ,  $B_2 = 1$ ,  $I = \Theta(\kappa)$  and  $T = \Theta(\log \frac{1}{\epsilon})$ . We have*

$$\mathbb{E}[\|\nabla p(\tilde{w}_T)\|_2^2] \leq \epsilon,$$

with the total sample complexity given by  $\mathcal{O}((\kappa + \frac{1}{\epsilon}) \log(\frac{1}{\epsilon}))$ .

Moreover, Algorithm 3 can be slightly modified to solve the minimization problem in eq. (7) in the finite-sum setting, in which

$$p(w) = \frac{1}{n} \sum_{i=1}^n P(w, \xi_i). \quad (8)$$

By replacing the large batch sample  $S_1$  used line 4 in Algorithm 3 with the full set of samples, we obtain the so-called SARAH algorithm Nguyen et al. (2017a). The following lemma characterizes the convergence result of SARAH to solve eq. (8).

**Lemma 7** (Nguyen et al. (2018), Corollary 2). *Consider Algorithm 3. Set  $\gamma = \Theta(\ell^{-1})$ ,  $B_2 = 1$ ,  $I = \Theta(\kappa)$  and  $T = \Theta(\log \frac{1}{\epsilon})$ . We have*

$$\mathbb{E}[\|\nabla p(\tilde{w}_T)\|_2^2] \leq \epsilon,$$

with the total sample complexity given by  $\mathcal{O}((\kappa + n) \log(\frac{1}{\epsilon}))$ .

### B.3 PROOF OF THEOREM 1

Throughout the paper, let  $n_t = \lceil t/q \rceil$  such that  $(n_t - 1)q \leq t \leq n_t q - 1$ . Without loss of generality, we assume  $\epsilon \leq 1$  since  $\epsilon$  is typically very small. Define  $\Delta_t = \mathbb{E}[\|\nabla_x f(x_t, y_t) - v_t\|_2^2] + \mathbb{E}[\|\nabla_y f(x_t, y_t) - u_t\|_2^2]$ ,  $\tilde{\Delta}_{t,k} = \mathbb{E}[\|\nabla_x f(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \tilde{v}_{t,k}\|_2^2] + \mathbb{E}[\|\nabla_y f(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \tilde{u}_{t,k}\|_2^2]$ , and  $\delta_t = \mathbb{E}[\|\nabla_y f(x_t, y_t)\|_2^2]$ .

We start our proof by a few supporting lemmas. The following lemma is a slightly modified version of Lemma 4 of Luo et al. (2020). The steps in the proof of Lemma 4 of Luo et al. (2020) does not yield their desired result.

**Lemma 8** (Modified version of Lemma 4 of Luo et al. (2020)). *Consider Algorithm 2. For all  $1 \leq t \leq m$ ,  $\beta \leq \frac{1}{2\ell}$  and  $S_2 \geq 2(\kappa + 1)\ell\beta$ . We have*

$$\mathbb{E}[\|\tilde{u}_{t,k}\|_2^2 | \mathcal{F}_{t,k}] \leq a \|\tilde{u}_{t,k-1}\|_2^2$$

where  $a = 1 - \frac{\mu\ell\beta}{\mu+\ell}$ .

Our Lemma 8 has the conditional number  $a = 1 - \frac{\mu\ell\beta}{\mu+\ell}$ , which is slightly larger than  $1 - \frac{2\mu\ell\beta}{\mu+\ell}$  given in Lemma 4 of Luo et al. (2020). The convergence analysis of SREDA in Luo et al. (2020) still holds but with  $a = 1 - \frac{\mu\ell\beta}{\mu+\ell}$ .

*Proof.* The update of Algorithm 2 implies that

$$\begin{aligned} & \mathbb{E}[\|\tilde{u}_{t,k}\|_2^2 | \mathcal{F}_{t,k}] \\ &= \|\tilde{u}_{t,k-1}\|_2^2 + 2\mathbb{E}[\langle \tilde{u}_{t,k-1}, \nabla_y G(\tilde{y}_{t,k}) - \nabla_y G(\tilde{y}_{t,k-1}) \rangle | \mathcal{F}_{t,k}] + \mathbb{E}[\|\nabla_y G(\tilde{y}_{t,k}) - \nabla_y G(\tilde{y}_{t,k-1})\|_2^2 | \mathcal{F}_{t,k}] \\ &= \|\tilde{u}_{t,k-1}\|_2^2 + \frac{2}{\beta} \mathbb{E}[\langle \tilde{y}_{t,k} - \tilde{y}_{t,k-1}, \nabla_y g(\tilde{y}_{t,k}) - \nabla_y g(\tilde{y}_{t,k-1}) \rangle] + \mathbb{E}[\|\nabla_y G(\tilde{y}_{t,k}) - \nabla_y G(\tilde{y}_{t,k-1})\|_2^2 | \mathcal{F}_{t,k}] \\ &\stackrel{(i)}{\leq} \|\tilde{u}_{t,k-1}\|_2^2 - \frac{2}{\beta} \left( \frac{\mu\ell}{\mu+\ell} \|\tilde{y}_{t,k} - \tilde{y}_{t,k-1}\|_2^2 + \frac{1}{\mu+\ell} \|\nabla_y g(\tilde{y}_{t,k}) - \nabla_y g(\tilde{y}_{t,k-1})\|_2^2 \right) \\ &\quad + \mathbb{E}[\|\nabla_y G(\tilde{y}_{t,k}) - \nabla_y G(\tilde{y}_{t,k-1})\|_2^2 | \mathcal{F}_{t,k}] \\ &\leq \left( 1 - \frac{2\mu\ell\beta}{\mu+\ell} \right) \|\tilde{u}_{t,k-1}\|_2^2 - \left( \frac{2}{\beta(\mu+\ell)} - 2 \right) \|\nabla_y g(\tilde{y}_{t,k}) - \nabla_y g(\tilde{y}_{t,k-1})\|_2^2 \\ &\quad + 2\mathbb{E}[\|\nabla_y G(\tilde{y}_{t,k}) - \nabla_y G(\tilde{y}_{t,k-1}) - [\nabla_y g(\tilde{y}_{t,k}) - \nabla_y g(\tilde{y}_{t,k-1})]\|_2^2 | \mathcal{F}_{t,k}] \end{aligned}$$

$$\begin{aligned}
&\stackrel{(ii)}{\leq} \left(1 - \frac{2\mu\ell\beta}{\mu + \ell}\right) \|\tilde{u}_{t,k-1}\|_2^2 + 2\mathbb{E}[\|\nabla_y G(\tilde{y}_{t,k}) - \nabla_y G(\tilde{y}_{t,k-1}) - [\nabla_y g(\tilde{y}_{t,k}) - \nabla_y g(\tilde{y}_{t,k-1})]\|_2^2 | \mathcal{F}_{t,k}] \\
&\stackrel{(iii)}{\leq} \left(1 - \frac{2\mu\ell\beta}{\mu + \ell}\right) \|\tilde{u}_{t,k-1}\|_2^2 + \frac{2\ell^2}{S_2} \|\tilde{y}_{t,k} - \tilde{y}_{t,k-1}\|_2^2 \\
&= \left(1 - \frac{2\mu\ell\beta}{\mu + \ell} + \frac{2\ell^2\beta^2}{S_2}\right) \|\tilde{u}_{t,k-1}\|_2^2 \\
&\stackrel{(iv)}{\leq} \left(1 - \frac{\mu\ell\beta}{\mu + \ell}\right) \|\tilde{u}_{t,k-1}\|_2^2, \tag{9}
\end{aligned}$$

where (i) follows from eq. (4) in Lemma 2, (ii) follows from the fact that  $\frac{1}{\beta(\mu+\ell)} - 1 \geq 0$  for all  $\beta \leq \frac{1}{2\ell}$ , (iii) follows from Lemma 5, and (iv) follows from the fact that  $S_2 \geq 2(\kappa + 1)\ell\beta$ .  $\square$

The following lemma implies the relationship between different estimation error terms, which can be obtained directly from the proof of (Luo et al., 2020, Lemma 5 in Section C).

**Lemma 9.** Suppose Assumption 1-4 hold. Let  $\beta \leq \frac{1}{\ell}$ . The following hold

$$\Delta_t \leq \tilde{\Delta}_{t-1,0} + \frac{\ell^2\beta^2}{S_2(1-a)} \mathbb{E}[\|\tilde{u}_{t-1,0}\|_2^2], \tag{10}$$

$$\begin{aligned}
\delta_t &\leq \frac{2}{\mu\beta(m+1)} (\mathbb{E}[\|\nabla_y f(x_t, y_{t-1}) - \nabla_y f(x_{t-1}, y_{t-1})\|_2^2] + \delta_{t-1}) \\
&\quad + \frac{\ell\beta}{2-\ell\beta} \mathbb{E}[\|\tilde{u}_{t-1,0}\|_2^2] + \tilde{\Delta}_{t-1,0}, \tag{11}
\end{aligned}$$

$$\tilde{\Delta}_{t,0} \leq \Delta_t + \frac{\ell^2}{S_2} \mathbb{E}[\|x_{t+1} - x_t\|_2^2], \tag{12}$$

$$\mathbb{E}[\|\tilde{u}_{t-1,0}\|_2^2] \leq 3(\tilde{\Delta}_{t-1,0} + \mathbb{E}[\|\nabla_y f(x_t, y_{t-1}) - \nabla_y f(x_{t-1}, y_{t-1})\|_2^2] + \delta_{t-1}). \tag{13}$$

*Proof.* eq. (10) can be obtained from the second inequality of eq. (23) in Luo et al. (2020) together with Lemma 8 as a correct version of Lemma 4 in Luo et al. (2020). eq. (11) can be obtained from the second inequality in the derivation of upper bound of " $\delta_{k+1}$ " in the page 22 of Luo et al. (2020). eq. (12) can be obtained from the second equality of eq. (22) in Luo et al. (2020). eq. (13) can be obtained from the first inequality of eq. (24) in Luo et al. (2020).  $\square$

We then provide the following lemma to characterize the induction relationships for  $\Delta_t$  and  $\delta_t$ .

**Lemma 10.** Suppose Assumption 1-4 hold. Let  $\beta \leq \frac{1}{\ell}$ . The following hold:

$$\Delta_t \leq \left(1 + \frac{3\ell^2\beta^2}{S_2(1-a)}\right) \Delta_{t-1} + \frac{\alpha^2\ell^2}{S_2} \left(1 + \frac{6\ell^2\beta^2}{1-a}\right) \mathbb{E}[\|v_{t-1}\|_2^2] + \frac{3\ell^2\beta^2}{S_2(1-a)} \delta_{t-1}, \tag{14}$$

$$\begin{aligned}
\delta_t &\leq \left(\frac{2}{\mu\beta(m+1)} + \frac{3\ell\beta}{2-\ell\beta}\right) \delta_{t-1} + \left(\frac{2\ell^2\alpha^2}{\mu\beta(m+1)} + \frac{6\ell^3\beta\alpha^2}{2-\ell\beta} + \ell^2\alpha^2\right) \mathbb{E}[\|v_{t-1}\|_2^2] \\
&\quad + \frac{2+2\ell\beta}{2-\ell\beta} \Delta_{t-1}. \tag{15}
\end{aligned}$$

*Proof.* To prove eq. (14), we proceed as follows:

$$\begin{aligned}
\Delta_t &\stackrel{(i)}{\leq} \tilde{\Delta}_{t-1,0} + \frac{\ell^2\beta^2}{S_2(1-a)} \mathbb{E}[\|\tilde{u}_{t-1,0}\|_2^2] \\
&\stackrel{(ii)}{\leq} \tilde{\Delta}_{t-1,0} + \frac{3\ell^2\beta^2}{S_2(1-a)} (\tilde{\Delta}_{t-1,0} + \mathbb{E}[\|\nabla_y f(x_t, y_{t-1}) - \nabla_y f(x_{t-1}, y_{t-1})\|_2^2] + \delta_{t-1}) \\
&\stackrel{(iii)}{\leq} \left(1 + \frac{3\ell^2\beta^2}{S_2(1-a)}\right) \tilde{\Delta}_{t-1,0} + \frac{3\ell^2\beta^2}{S_2(1-a)} (\ell^2\alpha^2 \mathbb{E}[\|v_{t-1}\|_2^2] + \delta_{t-1}) \\
&\stackrel{(iv)}{\leq} \left(1 + \frac{3\ell^2\beta^2}{S_2(1-a)}\right) \left(\Delta_{t-1} + \frac{\ell^2\alpha^2}{S_2} \mathbb{E}[\|v_{t-1}\|_2^2]\right) + \frac{3\ell^2\beta^2}{S_2(1-a)} (\ell^2\alpha^2 \mathbb{E}[\|v_{t-1}\|_2^2] + \delta_{t-1})
\end{aligned}$$

$$\leq \left(1 + \frac{3\ell^2\beta^2}{S_2(1-a)}\right) \Delta_{t-1} + \frac{\alpha^2\ell^2}{S_2} \left(1 + \frac{6\ell^2\beta^2}{1-a}\right) \mathbb{E}[\|v_{t-1}\|_2^2] + \frac{3\ell^2\beta^2}{S_2(1-a)}\delta_{t-1},$$

where (i) follows from eq. (10), (ii) follows from eq. (13), (iii) follows from Assumption 2, (iv) follows from eq. (12) and the fact that  $\|x_t - x_{t-1}\|_2 = \alpha \|v_{t-1}\|_2$ .

To prove eq. (15), we proceed as follows:

$$\begin{aligned} \delta_t &\stackrel{(i)}{\leq} \frac{2}{\mu\beta(m+1)} (\ell^2\alpha^2\mathbb{E}[\|v_{t-1}\|_2^2] + \delta_{t-1}) + \frac{\ell\beta}{2-\ell\beta} \mathbb{E}[\|\tilde{u}_{t-1,0}\|_2^2] + \tilde{\Delta}_{t-1,0} \\ &\stackrel{(ii)}{\leq} \frac{2}{\mu\beta(m+1)} (\ell^2\alpha^2\mathbb{E}[\|v_{t-1}\|_2^2] + \delta_{t-1}) + \frac{3\ell\beta}{2-\ell\beta} (\tilde{\Delta}_{t-1,0} + \ell^2\alpha^2\mathbb{E}[\|v_{t-1}\|_2^2] + \delta_{t-1}) + \tilde{\Delta}_{t-1,0} \\ &= \left(\frac{2}{\mu\beta(m+1)} + \frac{3\ell\beta}{2-\ell\beta}\right) \delta_{t-1} + \left(1 + \frac{3\ell\beta}{2-\ell\beta}\right) \tilde{\Delta}_{t-1,0} + \left(\frac{2\ell^2\alpha^2}{\mu\beta(m+1)} + \frac{3\ell^3\beta\alpha^2}{2-\ell\beta}\right) \mathbb{E}[\|v_{t-1}^2\|_2] \\ &\stackrel{(iii)}{\leq} \left(\frac{2}{\mu\beta(m+1)} + \frac{3\ell\beta}{2-\ell\beta}\right) \delta_{t-1} + \left(1 + \frac{3\ell\beta}{2-\ell\beta}\right) \left(\Delta_{t-1} + \frac{\ell^2\alpha^2}{S_2} \mathbb{E}[\|v_{t-1}\|_2^2]\right) \\ &\quad + \left(\frac{2\ell^2\alpha^2}{\mu\beta(m+1)} + \frac{3\ell^3\beta\alpha^2}{2-\ell\beta}\right) \mathbb{E}[\|v_{t-1}^2\|_2] \\ &\stackrel{(iv)}{\leq} \left(\frac{2}{\mu\beta(m+1)} + \frac{3\ell\beta}{2-\ell\beta}\right) \delta_{t-1} + \left(\frac{2\ell^2\alpha^2}{\mu\beta(m+1)} + \frac{6\ell^3\beta\alpha^2}{2-\ell\beta} + \ell^2\alpha^2\right) \mathbb{E}[\|v_{t-1}^2\|_2] \\ &\quad + \frac{2+2\ell\beta}{2-\ell\beta} \Delta_{t-1}, \end{aligned}$$

where (i) follows from eq. (11) and the fact that

$$\|\nabla_y f(x_t, y_{t-1}) - \nabla_y f(x_{t-1}, y_{t-1})\|_2 \leq \ell\alpha \|v_{t-1}\|_2, \quad (16)$$

(ii) follows from eq. (13), (iii) follows from eq. (12), and (iv) follows from the fact that  $S_2 \geq 1$ .  $\square$

We restate Theorem 1 as follows to include the specifics of the parameters.

**Theorem 3** (Restatement of Theorem 1). *Let Assumption 1-4 hold and apply SREDA-Boost in Algorithm 1 to solve the problem in eq. (1) with the following parameter choices:*

$$\begin{aligned} \zeta &= \frac{1}{\kappa}, \quad \alpha = \frac{1}{10(\kappa+1)\ell}, \quad \beta = \frac{2}{13\ell}, \quad q = \frac{2}{13(1+\kappa)} \frac{\kappa}{\epsilon}, \quad m = 52\kappa - 1, \\ S_1 &= \frac{9366\sigma^2\kappa^2}{\epsilon^2}, \quad S_2 = \frac{\kappa}{\epsilon}, \quad T = \max \left\{ \frac{3345\kappa}{\epsilon^2}, 6600(1+\kappa)\ell \frac{(\Phi(x_0) - \Phi^*)}{\epsilon^2} \right\}. \end{aligned}$$

Then, Algorithm 1 outputs  $\hat{x}$  that satisfies

$$\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2] \leq \epsilon$$

with at most  $\mathcal{O}(\kappa^3\epsilon^{-3})$  stochastic gradient evaluations.

**Proof Sketch of Theorem 1.** The proof of Theorem 1 consists of the following five steps.

**Step 1:** We establish the induction relationships for the tracking error and gradient estimation error upon one outer-loop update for SREDA-Boost. Namely, we develop the relationship between  $\delta_t$  and  $\delta_{t-1}$  as well as that between  $\Delta_t$  and  $\Delta_{t-1}$ .

**Step 2:** We provide the bounds on the inter-related accumulative errors  $\sum_{t=0}^{T-1} \Delta_t$  and  $\sum_{t=0}^{T-1} \delta_t$  over the entire execution of the algorithm.

**Step 3:** We decouple the bounds on  $\sum_{t=0}^{T-1} \Delta_t$  and  $\sum_{t=0}^{T-1} \delta_t$  in Step 2 from each other, and establish their separate relationships with the accumulative gradient estimators  $\sum_{i=0}^{T-1} \mathbb{E}[\|v_t\|_2^2]$ .

**Step 4:** We bound  $\sum_{i=0}^{T-1} \mathbb{E}[\|v_t\|_2^2]$ , and further cancel out the impact of  $\sum_{t=0}^{T-1} \Delta_t$  and  $\sum_{t=0}^{T-1} \delta_t$  by exploiting Step 3. **Step 5:** We establish the convergence bound on  $\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2]$  based on the bounds on its estimators  $\sum_{i=0}^{T-1} \mathbb{E}[\|v_t\|_2^2]$  and the two error bounds  $\sum_{t=0}^{T-1} \Delta_t$ , and  $\sum_{t=0}^{T-1} \delta_t$ .

*Proof of Theorem 1/Theorem 3.* By Lemma 3, the objective function  $\Phi$  is  $L$ -smooth, which implies that

$$\begin{aligned}
\Phi(x_{t+1}) &\leq \Phi(x_t) - \alpha \langle \nabla_x \Phi(x_t), v_t \rangle + \frac{L\alpha^2}{2} \|v_t\|_2^2 \\
&= \Phi(x_t) - \alpha \langle \nabla_x \Phi(x_t) - v_t, v_t \rangle - \alpha \|v_t\|_2^2 + \frac{L\alpha^2}{2} \|v_t\|_2^2 \\
&\stackrel{(i)}{\leq} \Phi(x_t) + \frac{\alpha}{2} \|\nabla_x \Phi(x_t) - v_t\|_2^2 + \frac{\alpha}{2} \|v_t\|_2^2 - \alpha \|v_t\|_2^2 + \frac{L\alpha^2}{2} \|v_t\|_2^2 \\
&\leq \Phi(x_t) + \frac{\alpha}{2} \|\nabla_x \Phi(x_t) - v_t\|_2^2 - \left( \frac{\alpha}{2} - \frac{L\alpha^2}{2} \right) \|v_t\|_2^2 \\
&\leq \Phi(x_t) + \alpha \|\nabla_x \Phi(x_t) - \nabla_x f(x_t, y_t)\|_2^2 + \alpha \|\nabla_x f(x_t, y_t) - v_t\|_2^2 - \left( \frac{\alpha}{2} - \frac{L\alpha^2}{2} \right) \|v_t\|_2^2 \\
&\stackrel{(ii)}{\leq} \Phi(x_t) + \alpha \kappa^2 \|\nabla_y f(x_t, y_t)\|_2^2 + \alpha \|\nabla_x f(x_t, y_t) - v_t\|_2^2 - \left( \frac{\alpha}{2} - \frac{L\alpha^2}{2} \right) \|v_t\|_2^2, \quad (17)
\end{aligned}$$

where (i) follows from the fact that  $(-1) \langle \nabla_x \Phi(x_t) - v_t, v_t \rangle \leq \frac{1}{2} \|\nabla_x \Phi(x_t) - v_t\|_2^2 + \frac{1}{2} \|v_t\|_2^2$ , and (ii) follows from the fact that

$$\begin{aligned}
\|\nabla_x \Phi(x_t) - \nabla_x f(x_t, y_t)\|_2^2 &= \|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t)\|_2^2 \leq \ell^2 \|y^*(x_t) - y_t\|_2^2 \\
&\stackrel{\text{eq. (5)}}{\leq} \frac{\ell^2}{\mu^2} \|\nabla_y f(x_t, y^*(x_t)) - \nabla_y f(x_t, y_t)\|_2^2 = \kappa^2 \|\nabla_y f(x_t, y_t)\|_2^2.
\end{aligned}$$

Taking expectation on both sides of eq. (17) yields

$$\begin{aligned}
\mathbb{E}[\Phi(x_{t+1})] &\leq \mathbb{E}[\Phi(x_t)] + \alpha \kappa^2 \mathbb{E}[\|\nabla_y f(x_t, y_t)\|_2^2] + \alpha \mathbb{E}[\|\nabla_x f(x_t, y_t) - v_t\|_2^2] - \left( \frac{\alpha}{2} - \frac{L\alpha^2}{2} \right) \mathbb{E}[\|v_t\|_2^2] \\
&\leq \mathbb{E}[\Phi(x_t)] + \alpha \kappa^2 \delta_t + \alpha \Delta_t - \left( \frac{\alpha}{2} - \frac{L\alpha^2}{2} \right) \mathbb{E}[\|v_t\|_2^2]. \quad (18)
\end{aligned}$$

Rearranging eq. (18) and summing over  $t = \{0, \dots, T-1\}$  yield

$$\left( \frac{\alpha}{2} - \frac{L\alpha^2}{2} \right) \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \leq \Phi(x_0) - \mathbb{E}[\Phi(x_T)] + \alpha \kappa^2 \sum_{t=0}^{T-1} \delta_t + \alpha \sum_{t=0}^{T-1} \Delta_t. \quad (19)$$

Then, we proceed the proof in the following five steps.

**Step 1.** We establish the induction relationships for the tracking error and gradient estimation error upon one outer-loop update for SREDA-Boost. Namely, we develop the relationship between  $\delta_t$  and  $\delta_{t-1}$  as well as that between  $\Delta_t$  and  $\Delta_{t-1}$ , which are captured in Lemma 10.

**Step 2.** Based on Step 1, we provide the bounds on the inter-related accumulative errors  $\sum_{t=0}^{T-1} \Delta_t$  and  $\sum_{t=0}^{T-1} \delta_t$  over the entire execution of the algorithm.

We first consider  $\sum_{t=0}^{T-1} \Delta_t$ . For any  $(n_T - 1)q \leq t' < p < T - 1$ , we apply eq. (14) recursively to obtain the following inequality

$$\begin{aligned}
\Delta_t &\leq \left( 1 + \frac{3\ell^2\beta^2}{S_2(1-a)} \right) \Delta_{t-1} + \frac{\alpha^2\ell^2}{S_2} \left( 1 + \frac{6\ell^2\beta^2}{1-a} \right) \mathbb{E}[\|v_{t-1}\|_2^2] + \frac{3\ell^2\beta^2}{S_2(1-a)} \delta_{t-1} \\
&\leq \left( 1 + \frac{3\ell^2\beta^2}{S_2(1-a)} \right)^{t-t'} \Delta_{t'} + \frac{\alpha^2\ell^2}{S_2} \left( 1 + \frac{6\ell^2\beta^2}{1-a} \right) \sum_{p=t'}^{t-1} \left( 1 + \frac{3\ell^2\beta^2}{S_2(1-a)} \right)^{p-t'} \mathbb{E}[\|v_t\|_2^2] \\
&\quad + \frac{3\ell^2\beta^2}{S_2(1-a)} \sum_{p=t'}^{t-1} \left( 1 + \frac{3\ell^2\beta^2}{S_2(1-a)} \right)^{p-t'} \delta_t \\
&\stackrel{(i)}{\leq} 2\Delta_{t'} + \frac{2\alpha^2\ell^2}{S_2} \left( 1 + \frac{6\ell^2\beta^2}{1-a} \right) \sum_{p=t'}^{t-1} \mathbb{E}[\|v_t\|_2^2] + \frac{6\ell^2\beta^2}{S_2(1-a)} \sum_{p=t'}^{t-1} \delta_t, \quad (20)
\end{aligned}$$

where (i) follows from the fact that

$$\begin{aligned} \left(1 + \frac{3\ell^2\beta^2}{S_2(1-a)}\right)^{p-t'} &\leq \left(1 + \frac{3\ell^2\beta^2}{S_2(1-a)}\right)^q \stackrel{(ii)}{\leq} 1 + \frac{\frac{3q\ell^2\beta^2}{S_2(1-a)}}{1 - \frac{3(q-1)\ell^2\beta^2}{S_2(1-a)}} \leq 1 + \frac{\frac{3q\ell^2\beta^2}{S_2(1-a)}}{1 - \frac{3q\ell^2\beta^2}{S_2(1-a)}} \\ &\stackrel{(iii)}{=} 1 + \frac{3\ell^2\beta^2}{1 - 3\ell^2\beta^2} \stackrel{(iv)}{<} 2, \end{aligned}$$

where (ii) follows from Bernoulli's inequality [Li & Yeh \(2013\)](#)

$$(1+c)^r \leq 1 + \frac{rc}{1-(r-1)c} \quad \text{for } c \in \left[-1, \frac{1}{r-1}\right), r > 1, \quad (21)$$

(iii) follows from the fact that  $q = (1-a)S_2$  and (iv) follows from the fact that  $\beta = \frac{2}{13\ell}$ .

Letting  $t' = (n_T - 1)q$  and taking summation of eq. (20) over  $t = \{(n_T - 1)q, \dots, T-1\}$  yield

$$\begin{aligned} \sum_{t=(n_T-1)q}^{T-1} \Delta_t &\leq 2(T - (n_T - 1)q)\Delta_{(n_T-1)q} + \frac{2\alpha^2\ell^2}{S_2} \left(1 + \frac{6\ell^2\beta^2}{1-a}\right) \sum_{t=(n_T-1)q}^{T-1} \sum_{p=(n_T-1)q}^{t-1} \mathbb{E}[\|v_t\|_2^2] \\ &\quad + \frac{6\ell^2\beta^2}{S_2(1-a)} \sum_{t=(n_T-1)q}^{T-1} \sum_{p=(n_T-1)q}^{t-1} \delta_t \\ &\stackrel{(i)}{\leq} 2(T - (n_T - 1)q)\frac{\sigma^2}{S_1} + \frac{2\alpha^2\ell^2q}{S_2} \left(1 + \frac{6\ell^2\beta^2}{1-a}\right) \sum_{t=(n_T-1)q}^{T-2} \mathbb{E}[\|v_t\|_2^2] \\ &\quad + \frac{6\ell^2\beta^2q}{S_2(1-a)} \sum_{t=(n_T-1)q}^{T-2} \delta_t, \end{aligned} \quad (22)$$

where (i) follows from the fact that  $\Delta_{(n_T-n)q} \leq \frac{\sigma^2}{S_1}$  for all  $n \leq n_T$  (following [Lemma 4](#)),

$$\sum_{t=(n_T-1)q}^{T-1} \sum_{p=(n_T-1)q}^{t-1} \mathbb{E}[\|v_t\|_2^2] \leq q \sum_{t=(n_T-1)q}^{T-2} \mathbb{E}[\|v_t\|_2^2],$$

and

$$\sum_{t=(n_T-1)q}^{T-1} \sum_{p=(n_T-1)q}^{t-1} \delta_t \leq q \sum_{t=(n_T-1)q}^{T-2} \delta_t.$$

Applying steps similar to those in eq. (22) for iterations over  $p = \{(n_T - n_t)q, \dots, (n_T - n_t + 1)q - 1\}$  (where  $n_t$  is an integer that satisfies  $2 \leq n_t < n_T$ ) yields

$$\begin{aligned} \sum_{t=(n_T-n_t)q}^{(n_T-n_t+1)q-1} \Delta_t &\leq \frac{2\sigma^2q}{S_1} + \frac{2\alpha^2\ell^2q}{S_2} \left(1 + \frac{6\ell^2\beta^2}{1-a}\right) \sum_{t=(n_T-n_t)q}^{(n_T-n_t+1)q-1} \mathbb{E}[\|v_t\|_2^2] \\ &\quad + \frac{6\ell^2\beta^2q}{S_2(1-a)} \sum_{t=(n_T-n_t)q}^{(n_T-n_t+1)q-1} \delta_t. \end{aligned} \quad (23)$$

Taking summation of eq. (23) over  $n_t = \{2, \dots, n_T\}$  and combining with eq. (22) yield

$$\begin{aligned} \sum_{t=0}^{T-1} \Delta_t &\leq \frac{2\sigma^2T}{S_1} + \frac{2\alpha^2\ell^2q}{S_2} \left(1 + \frac{6\ell^2\beta^2}{1-a}\right) \sum_{t=0}^{T-2} \mathbb{E}[\|v_t\|_2^2] + \frac{6\ell^2\beta^2q}{S_2(1-a)} \sum_{t=0}^{T-2} \delta_t \\ &\leq \frac{2\sigma^2T}{S_1} + 4\alpha^2\ell^2 \sum_{t=0}^{T-2} \mathbb{E}[\|v_t\|_2^2] + \frac{1}{5} \sum_{t=0}^{T-2} \delta_t. \end{aligned} \quad (24)$$

Then we consider the upper bound on  $\sum_{t=0}^{T-1} \delta_t$ . Since  $m = \frac{8}{\mu\beta} - 1$  and  $\beta = \frac{2}{13\ell}$ , eq. (15) implies

$$\delta_t \leq \frac{1}{2}\delta_{t-1} + \frac{7}{4}\ell^2\alpha^2\mathbb{E}[\|v_{t-1}\|_2^2] + \frac{5}{4}\Delta_{t-1}, \quad (25)$$

for all  $t \geq 1$ . Applying eq. (25) recursively yields

$$\delta_t \leq \frac{1}{2^t}\delta_0 + \frac{7}{4}\ell^2\alpha^2\sum_{t=0}^{t-1}\frac{1}{2^t}\mathbb{E}[\|v_t\|_2^2] + \frac{5}{4}\sum_{t=0}^{t-1}\frac{1}{2^t}\Delta_t. \quad (26)$$

Taking the summation of eq. (26) over  $t = \{0, 1, \dots, T-1\}$  yields

$$\begin{aligned} \sum_{t=0}^{T-1} \delta_t &\leq \delta_0 \sum_{t=0}^{T-1} \frac{1}{2^t} + \frac{7}{4}\ell^2\alpha^2 \sum_{p=0}^{T-1} \sum_{t=0}^{p-1} \frac{1}{2^t} \mathbb{E}[\|v_t\|_2^2] + \frac{5}{4} \sum_{p=0}^{T-1} \sum_{t=0}^{p-1} \frac{1}{2^t} \Delta_t \\ &\leq 2\delta_0 + \frac{7}{2}\ell^2\alpha^2 \sum_{t=0}^{T-2} \mathbb{E}[\|v_t\|_2^2] + \frac{5}{2} \sum_{t=0}^{T-2} \Delta_t. \end{aligned} \quad (27)$$

**Step 3.** We decouple the bounds on  $\sum_{t=0}^{T-1} \Delta_t$  and  $\sum_{t=0}^{T-1} \delta_t$  in Step 2 from each other, and establish their separate relationships with the accumulative gradient estimators  $\sum_{i=0}^{T-1} \mathbb{E}[\|v_t\|_2^2]$ .

Substituting eq. (27) into eq. (24) yields

$$\sum_{t=0}^{T-1} \Delta_t \leq \frac{2\sigma^2 T}{S_1} + \frac{2}{5}\delta_0 + 5\alpha^2\ell^2 \sum_{t=0}^{T-2} \mathbb{E}[\|v_t\|_2^2] + \frac{1}{2} \sum_{t=0}^{T-2} \Delta_t,$$

which implies

$$\sum_{t=0}^{T-1} \Delta_t \leq \frac{4\sigma^2 T}{S_1} + \frac{4}{5}\delta_0 + 10\alpha^2\ell^2 \sum_{t=0}^{T-2} \mathbb{E}[\|v_t\|_2^2]. \quad (28)$$

Substituting eq. (28) into eq. (27) yields

$$\sum_{t=0}^{T-1} \delta_t \leq \frac{10\sigma^2 T}{S_1} + 4\delta_0 + 30\alpha^2\ell^2 \sum_{t=0}^{T-2} \mathbb{E}[\|v_t\|_2^2]. \quad (29)$$

**Step 4.** We bound  $\sum_{i=0}^{T-1} \mathbb{E}[\|v_i\|_2^2]$ , and further cancel out the impact of  $\sum_{t=0}^{T-1} \Delta_t$  and  $\sum_{t=0}^{T-1} \delta_t$  by exploiting Step 3.

Substituting eq. (28) and eq. (29) into eq. (19) yields

$$\begin{aligned} \left(\frac{\alpha}{2} - \frac{L\alpha^2}{2}\right) \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] &\leq \Phi(x_0) - \mathbb{E}[\Phi(x_T)] + (10\kappa^2 + 4) \frac{\alpha\sigma^2 T}{S_1} + \left(4\kappa^2 + \frac{4}{5}\right) \alpha\delta_0 \\ &\quad + 10\alpha^3\ell^2 (3\kappa^2 + 1) \sum_{t=0}^{T-2} \mathbb{E}[\|v_t\|_2^2] \\ &\stackrel{(i)}{\leq} \Phi(x_0) - \mathbb{E}[\Phi(x_T)] + \frac{14\alpha\kappa^2\sigma^2 T}{S_1} + 5\kappa^2\alpha\delta_0 + 40\alpha^3 L^2 \sum_{t=0}^{T-2} \mathbb{E}[\|v_t\|_2^2], \end{aligned} \quad (30)$$

where (i) follows from the fact that  $L = (1 + \kappa)\ell$  and  $\kappa > 1$ . Rearranging eq. (30), we have

$$\left(\frac{\alpha}{2} - \frac{L\alpha^2}{2} - 40L^2\alpha^3\right) \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \leq \Phi(x_0) - \mathbb{E}[\Phi(x_T)] + \frac{14\alpha\kappa^2\sigma^2 T}{S_1} + 5\kappa^2\alpha\delta_0. \quad (31)$$

Since  $\alpha = \frac{1}{10L}$ , we obtain

$$\frac{\alpha}{2} - \frac{L\alpha^2}{2} - 40L^2\alpha^3 = \frac{1}{200L}. \quad (32)$$



Substituting eq. (32) into eq. (31) and applying Assumption 1 yield

$$\sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \leq 200L(\Phi(x_0) - \Phi^*) + \frac{280\kappa^2\sigma^2T}{S_1} + 100\kappa^2\delta_0. \quad (33)$$

**Step 5.** We establish the convergence bound on  $\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2]$  based on the bounds on its estimators  $\sum_{i=0}^{T-1} \mathbb{E}[\|v_t\|_2^2]$  and the two error bounds  $\sum_{t=0}^{T-1} \Delta_t$ , and  $\sum_{t=0}^{T-1} \delta_t$ .

Observe that

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\Phi(x_t)\|_2^2] &\leq \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\Phi(x_t) - \nabla_x f(x_t, y_t) + \nabla_x f(x_t, y_t) - v_t + v_t\|_2^2] \\ &\leq 3 \sum_{t=0}^{T-1} \left( \mathbb{E}[\|\nabla\Phi(x_t) - \nabla_x f(x_t, y_t)\|_2^2] + \mathbb{E}[\|\nabla_x f(x_t, y_t) - v_t\|_2^2] + \mathbb{E}[\|v_t\|_2^2] \right) \\ &\leq 3 \sum_{t=0}^{T-1} \left( \kappa^2 \mathbb{E}[\|\nabla_y f(x_t, y_t)\|_2^2] + \mathbb{E}[\|\nabla_x f(x_t, y_t) - v_t\|_2^2] + \mathbb{E}[\|v_t\|_2^2] \right) \\ &\leq 3\kappa^2 \sum_{t=0}^{T-1} \delta_t + 3 \sum_{t=0}^{T-1} \Delta_t + 3 \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2]. \end{aligned} \quad (34)$$

Substituting eq. (28), eq. (29) and eq. (33) into eq. (34), and using the fact that  $\kappa \geq 1$  yield

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\Phi(x_t)\|_2^2] &\leq 42\kappa^2 \frac{\sigma^2 T}{S_1} + 15\kappa^2\delta_0 + 11 \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \\ &\leq 2200L(\Phi(x_0) - \Phi^*) + \frac{3122\kappa^2\sigma^2T}{S_1} + 1115\kappa^2\delta_0. \end{aligned} \quad (35)$$

Recall  $L = (1 + \kappa)\ell$ . Then, eq. (35) implies that

$$\begin{aligned} \mathbb{E}[\|\nabla\Phi(\hat{x})\|_2^2] &= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\Phi(x_t)\|_2^2] \\ &\leq 2200(1 + \kappa)\ell \frac{\Phi(x_0) - \Phi^*}{T} + \frac{3122\kappa^2\sigma^2}{S_1} + 1115 \frac{\kappa^2\delta_0}{T}. \end{aligned} \quad (36)$$

If we let  $\delta_0 = \frac{1}{\kappa}$ ,  $T = \max\{\frac{3345\kappa}{\epsilon^2}, 6600(1 + \kappa)\ell \frac{(\Phi(x_0) - \Phi^*)}{\epsilon^2}\}$ ,  $S_1 = \frac{9366\sigma^2\kappa^2}{\epsilon^2}$ ,  $S_2 = \frac{\kappa}{\epsilon}$ , and  $q = (1 - a)S_2$ , then we have

$$\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2] \leq \sqrt{\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2^2]} \leq \epsilon.$$

We define  $T_0$  as the sample complexity of iSARAH to achieve the accuracy  $\mathbb{E}[\|\nabla_y f(x_0, y_0)\|_2^2] \leq \frac{1}{\kappa}$ . Lemma 6 implies that  $T_0 = \mathcal{O}(\kappa \log(\kappa))$ . Then, the total sample complexity is given by

$$\begin{aligned} T \cdot S_2 \cdot m + \left\lceil \frac{T}{q} \right\rceil \cdot S_1 + T_0 &\leq \Theta\left(\frac{\kappa}{\epsilon^2} \cdot \frac{\kappa}{\epsilon} \cdot \kappa\right) + \Theta\left(\frac{\kappa}{\epsilon} \cdot \frac{\kappa^2}{\epsilon^2}\right) + \mathcal{O}(\kappa \log(\kappa)) \\ &= \mathcal{O}\left(\frac{\kappa^3}{\epsilon^3}\right), \end{aligned}$$

which completes the proof.  $\square$

#### B.4 PROOF OF COROLLARY 1

In the finite-sum case, recall that we have

$$f(x, y) \triangleq \frac{1}{n} \sum_{i=1}^n F(x, y; \xi_i).$$

Here we modify Algorithm 5 by replacing the large batch sample used in line 6 of Algorithm 1 with the full gradient and using SARAH [Nguyen et al. \(2017a\)](#) as initialization.

**Case 1:**  $n > \kappa^2$

In the finite-sum case, due to the utilization of the full gradient every  $q$  steps, we have  $S_1 = n$  and  $\Delta_{(n_T-n)q} = 0$  for all  $n \leq n_T$ . Then following steps similar to those from eq. (22) to eq. (36), we obtain

$$\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2^2] \leq 2200(1+\kappa)\ell \frac{\Phi(x_0) - \Phi^*}{T} + 1115 \frac{\kappa^2 \delta_0}{T}.$$

If we let  $\delta_0 = \frac{1}{\kappa}$ ,  $T = \max\{\frac{2230\kappa}{\epsilon^2}, 4400(1+\kappa)\ell \frac{(\Phi(x_0) - \Phi^*)}{\epsilon^2}\}$ ,  $S_2 = \sqrt{n}$ , and  $q = \lceil (1-a)S_2 \rceil$ , then we have

$$\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2] \leq \sqrt{\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2^2]} \leq \epsilon.$$

We define  $T_0$  as the sample complexity of SARAH to achieve the accuracy  $\mathbb{E}[\|\nabla_y f(x_0, y_0)\|_2^2] \leq \frac{1}{\kappa}$ . Lemma 7 implies that  $T_0 = \mathcal{O}((n+\kappa)\log(\kappa))$ . Then, the total sample complexity is given by

$$\begin{aligned} T \cdot S_2 \cdot m + \left\lceil \frac{T}{q} \right\rceil \cdot S_1 + T_0 &\leq \Theta\left(\frac{\kappa}{\epsilon^2} \cdot \sqrt{n} \cdot \kappa\right) + \Theta\left(\left\lceil \frac{\kappa^2}{\epsilon^2 \sqrt{n}} \right\rceil \cdot n\right) + \mathcal{O}((n+\kappa)\log(\kappa)) \\ &= \mathcal{O}(\kappa^2 \sqrt{n} \epsilon^{-2} + n) + \mathcal{O}((n+\kappa)\log(\kappa)). \end{aligned}$$

**Case 2:**  $n \leq \kappa^2$

In this case, we let  $q = 1$  and  $S_2 = 1$ . Then, we have  $\Delta_t = 0$  for all  $0 \leq t \leq T-1$ . Since the analysis of  $\delta_t$  does not depend on the value of  $S_2$ , eq. (25) still holds, which implies

$$\delta_t \leq \frac{1}{2}\delta_{t-1} + \frac{7}{4}\ell^2 \alpha^2 \mathbb{E}[\|v_{t-1}\|_2^2].$$

Following steps similar to those from eq. (25)-27 yields

$$\sum_{t=0}^{T-1} \delta_t \leq 2\delta_0 + \frac{7}{2}\ell^2 \alpha^2 \sum_{t=0}^{T-2} \mathbb{E}[\|v_t\|_2^2]. \quad (37)$$

Substituting eq. (37) into eq. (18) yields

$$\begin{aligned} \left(\frac{\alpha}{2} - \frac{L\alpha^2}{2}\right) \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] &\leq \Phi(x_0) - \mathbb{E}[\Phi(x_T)] + 2\alpha\kappa^2\delta_0 + \frac{7}{2}\ell^2 \kappa^2 \alpha^3 \sum_{t=0}^{T-2} \mathbb{E}[\|v_t\|_2^2] \\ &\stackrel{(i)}{\leq} \Phi(x_0) - \mathbb{E}[\Phi(x_T)] + 2\alpha\kappa^2\delta_0 + \frac{7}{2}L^2 \alpha^3 \sum_{t=0}^{T-2} \mathbb{E}[\|v_t\|_2^2], \end{aligned} \quad (38)$$

where (i) follows from the fact that  $L = (1+\kappa)\ell$ . Rearranging eq. (38), we have

$$\left(\frac{\alpha}{2} - \frac{L\alpha^2}{2} - \frac{7}{2}L^2 \alpha^3\right) \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \leq \Phi(x_0) - \mathbb{E}[\Phi(x_T)] + 2\alpha\kappa^2\delta_0. \quad (39)$$

Let  $\alpha = \frac{1}{4L}$ . We have

$$\frac{\alpha}{2} - \frac{L\alpha^2}{2} - \frac{7}{2}L^2 \alpha^3 = \frac{5}{128L}. \quad (40)$$

Combining eq. (40) and eq. (39) and using Assumption 1 yield

$$\sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \leq 26L(\Phi(x_0) - \Phi^*) + 14\kappa^2\delta_0. \quad (41)$$

Recalling eq. (34), we have

$$\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\Phi(x_t)\|_2^2] \leq 3\kappa^2 \sum_{t=0}^{T-1} \delta_t + 3 \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2]. \quad (42)$$

Substituting eq. (37) and eq. (41) into eq. (42) yields

$$\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\Phi(x_t)\|_2^2] \leq 6\kappa^2\delta_0 + 4 \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \leq 62\kappa^2\delta_0 + 104L(\Phi(x_0) - \Phi^*). \quad (43)$$

Recall that  $L = (1 + \kappa)\ell$ . Then eq. (43) implies that

$$\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2^2] = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\Phi(x_t)\|_2^2] \leq \frac{62\kappa^2\delta_0}{T} + \frac{104(\kappa + 1)\ell(\Phi(x_0) - \Phi^*)}{T}. \quad (44)$$

Let  $\delta_0 \leq \frac{1}{\kappa}$  and  $T = \max\{\frac{124\kappa}{\epsilon^2}, \frac{208(\kappa+1)(\Phi(x_0)-\Phi^*)}{\epsilon}\}$ . Then we have

$$\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2] \leq \sqrt{\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2^2]} \leq \epsilon.$$

The total sample complexity is given by

$$\begin{aligned} T \cdot S_2 \cdot m + \left\lceil \frac{T}{q} \right\rceil \cdot S_1 + T_0 &\leq \Theta\left(\frac{\kappa}{\epsilon^2} \cdot 1 \cdot \kappa\right) + \Theta\left(\left\lceil \frac{\kappa}{\epsilon^2} \right\rceil \cdot n\right) + \mathcal{O}((n + \kappa) \log(\kappa)) \\ &= \mathcal{O}((\kappa^2 + \kappa n)\epsilon^{-2}). \end{aligned}$$

## C CONVERGENCE ANALYSIS OF ZERO-TH-ORDER SREDA-BOOST

### C.1 ZO-SREDA-BOOST ALGORITHM

---

#### Algorithm 4 ZO-SREDA-Boost

---

```

1: Input:  $x_0$ , initial accuracy  $\zeta$ , learning rate  $\alpha = \Theta(\frac{1}{\kappa\ell})$ ,  $\beta = \Theta(\frac{1}{\ell})$ , batch size  $S_1, S_2$  and periods  $q, m$ .
2: Initialization:  $y_0 = \text{ZO-iSARAH}(-f(x_0, \cdot), \zeta, \mu_2)$  (detailed in Algorithm 6 in Appendix C.4)
3: for  $t = 0, 1, \dots, T - 1$  do
4:   if  $\text{mod}(k, q) = 0$  then draw  $S_1$  samples  $\{\xi_1, \dots, \xi_{S_1}\}$ 
5:      $v_t = \frac{1}{S_1} \sum_{i=1}^{S_1} \sum_{j=1}^{d_1} \frac{F(x_t + \delta e_j, y_t, \xi_i) - F(x_t - \delta e_j, y_t, \xi_i)}{2\delta} e_j$ 
6:      $u_t = \frac{1}{S_1} \sum_{i=1}^{S_1} \sum_{j=1}^{d_2} \frac{F(x_t, y_t + \delta e_j, \xi_i) - F(x_t, y_t - \delta e_j, \xi_i)}{2\delta} e_j$ 
7:     where  $e_j$  denotes the vector with  $j$ -th natural unit basis vector.
8:   else
9:      $v_t = \tilde{v}_{t-1, \tilde{m}_{t-1}}, u_t = \tilde{u}_{t-1, \tilde{m}_{t-1}}$ 
10:  end if
11:   $x_{t+1} = x_t - \alpha v_t$ 
12:   $y_{t+1} = \text{ZO-ConcaveMaximizer}(t, m, S_{2,x}, S_{2,y})$ 
13: end for
14: Output:  $\hat{x}$  chosen uniformly at random from  $\{x_t\}_{t=0}^{T-1}$ 

```

---



---

#### Algorithm 5 ZO-ConcaveMaximizer( $t, m, S_{2,x}, S_{2,y}$ )

---

```

1: Initialization:  $\tilde{x}_{t,-1} = x_t, \tilde{y}_{t,-1} = y_t, \tilde{x}_{t,0} = x_{t+1}, \tilde{y}_{t,0} = y_t, \tilde{v}_{t,-1} = v_t, \tilde{u}_{t,-1} = u_t$ 
2: Draw minibatch sample  $\mathcal{M}_x = \{\xi_1, \dots, \xi_{S_{2,x}}\}$ ,  $\mathcal{M}_{1,x} = \{\nu_1, \dots, \nu_{S_{2,x}}\}$  and  $\mathcal{M}_{2,x} = \{\omega_1, \dots, \omega_{S_{2,x}}\}$ , and  $\mathcal{M}_y = \{\xi_1, \dots, \xi_{S_{2,y}}\}$ ,  $\mathcal{M}_{1,y} = \{\nu_1, \dots, \nu_{S_{2,y}}\}$  and  $\mathcal{M}_{2,y} = \{\omega_1, \dots, \omega_{S_{2,y}}\}$ 
3:  $\tilde{v}_{t,0} = \tilde{v}_{t,-1} + G(\tilde{x}_{t,0}, \tilde{y}_{t,0}, \nu_{\mathcal{M}_{1,x}}, \xi_{\mathcal{M}_x}) - G(\tilde{x}_{t,-1}, \tilde{y}_{t,-1}, \nu_{\mathcal{M}_{1,x}}, \xi_{\mathcal{M}_x})$ 
4:  $\tilde{u}_{t,0} = \tilde{u}_{t,-1} + H(\tilde{x}_{t,0}, \tilde{y}_{t,0}, \omega_{\mathcal{M}_{2,y}}, \xi_{\mathcal{M}_y}) - H(\tilde{x}_{t,-1}, \tilde{y}_{t,-1}, \omega_{\mathcal{M}_{2,y}}, \xi_{\mathcal{M}_y})$ 
5:  $\tilde{x}_{t,1} = \tilde{x}_{t,0}$ 
6:  $\tilde{y}_{t,1} = \tilde{y}_{t,0} + \beta \tilde{u}_{t,0}$ 
7: for  $k = 1, 2, \dots, m + 1$  do
8:   Draw minibatch sample  $\mathcal{M}_x = \{\xi_1, \dots, \xi_{S_{2,x}}\}$ ,  $\mathcal{M}_{1,x} = \{\nu_1, \dots, \nu_{S_{2,x}}\}$  and  $\mathcal{M}_{2,x} = \{\omega_1, \dots, \omega_{S_{2,x}}\}$ , and  $\mathcal{M}_y = \{\xi_1, \dots, \xi_{S_{2,y}}\}$ ,  $\mathcal{M}_{1,y} = \{\nu_1, \dots, \nu_{S_{2,y}}\}$  and  $\mathcal{M}_{2,y} = \{\omega_1, \dots, \omega_{S_{2,y}}\}$ 
9:    $\tilde{v}_{t,k} = \tilde{v}_{t,k-1} + G_{\mu_1}(\tilde{x}_{t,k}, \tilde{y}_{t,k}, \nu_{\mathcal{M}_{1,x}}, \xi_{\mathcal{M}_x}) - G_{\mu_1}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}, \nu_{\mathcal{M}_{1,x}}, \xi_{\mathcal{M}_1})$ 
10:   $\tilde{u}_{t,k} = \tilde{u}_{t,k-1} + H_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}, \omega_{\mathcal{M}_{2,y}}, \xi_{\mathcal{M}_y}) - H_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}, \omega_{\mathcal{M}_{2,y}}, \xi_{\mathcal{M}_y})$ 
11:   $\tilde{x}_{t,k+1} = \tilde{x}_{t,k}$ 
12:   $\tilde{y}_{t,k+1} = \tilde{y}_{t,k} + \beta \tilde{u}_{t,k}$ 
13: end for
Output:  $y_{t+1} = \tilde{y}_{t, \tilde{m}_t}$  with  $\tilde{m}_t$  chosen uniformly at random from  $\{0, 1, \dots, m\}$ 

```

---

## C.2 PRELIMINARIES

Consider a function  $h(\cdot): \mathbb{R}^d \rightarrow \mathbb{R}$ . Let  $\nu$  be a  $d$ -dimensional standard Gaussian random vector and  $\mu > 0$  be the smoothing parameter. Then a smooth approximation of  $h(\cdot)$  is defined as  $h_\tau(x) = \mathbb{E}_\nu[h(x + \tau\nu)]$ . We have the following lemmas.

**Lemma 11** (Nesterov & Spokoiny (2017), Section 2). *If  $h(\cdot)$  is convex, then  $h_\mu(\cdot)$  is also a convex function.*

**Lemma 12** (Ghadimi & Lan (2013), Section 3.1). *If  $h(\cdot)$  has  $\ell$ -Lipschitz gradient, then  $h_\mu(\cdot)$  also has  $\ell$ -Lipschitz gradient.*

**Lemma 13** (Nesterov & Spokoiny (2017), Theorem 1). *If  $h(\cdot)$  has  $\ell$ -Lipschitz gradient, then for all  $x \in \mathbb{R}^d$ , we have  $|h(x) - h_\tau(x)| \leq \frac{\tau^2}{2}\ell d$ .*

**Lemma 14** (Nesterov & Spokoiny (2017), Lemma 3). *If  $h(\cdot)$  has  $\ell$ -Lipschitz gradient, then  $\|\nabla_x h_\tau(x) - \nabla_x h(x)\|_2^2 \leq \frac{\tau^2}{4}\ell^2(d+3)^3$ .*

**Lemma 15.** *Suppose Assumption 2 and Assumption 4 hold. Suppose  $\text{mod}(t, q) = 0$ , and let  $\epsilon(S_1, \delta) = \mathbb{E}[\|v_t - \nabla_x f_{\mu_1}(x_t, y_t)\|_2^2] + \mathbb{E}[\|u_t - \nabla_y f_{\mu_2}(x_t, y_t)\|_2^2]$ . Then, we have*

$$\epsilon(S_1, \delta) \leq \frac{(d_1 + d_2)\ell^2\delta^2}{2} + \frac{4\sigma^2}{S_1} + \frac{\mu_1^2}{2}\ell^2(d_1 + 3)^3 + \frac{\mu_2^2}{2}\ell^2(d_2 + 3)^3.$$

*Proof.* B.56 and B.57 in Fang et al. (2018) imply that

$$\mathbb{E}[\|v_t - \nabla_x f(x_t, y_t)\|_2^2] \leq \frac{d_1\ell^2\delta^2}{2} + \frac{2\sigma^2}{S_1}, \quad (45)$$

and

$$\mathbb{E}[\|u_t - \nabla_y f(x_t, y_t)\|_2^2] \leq \frac{d_2\ell^2\delta^2}{2} + \frac{2\sigma^2}{S_1}. \quad (46)$$

Then we proceed as follows:

$$\begin{aligned} & \mathbb{E}[\|v_t - \nabla_x f_{\mu_1}(x_t, y_t)\|_2^2] + \mathbb{E}[\|u_t - \nabla_y f_{\mu_2}(x_t, y_t)\|_2^2] \\ & \leq 2\mathbb{E}[\|v_t - \nabla_x f(x_t, y_t)\|_2^2] + 2\mathbb{E}[\|u_t - \nabla_y f(x_t, y_t)\|_2^2] \\ & \quad + 2\mathbb{E}[\|\nabla_x f_{\mu_1}(x_t, y_t) - \nabla_x f(x_t, y_t)\|_2^2] + 2\mathbb{E}[\|\nabla_y f_{\mu_2}(x_t, y_t) - \nabla_y f(x_t, y_t)\|_2^2] \\ & \stackrel{(i)}{\leq} 2\mathbb{E}[\|v_t - \nabla_x f(x_t, y_t)\|_2^2] + 2\mathbb{E}[\|u_t - \nabla_y f(x_t, y_t)\|_2^2] + \frac{\mu_1^2}{2}\ell^2(d_1 + 3)^3 + \frac{\mu_2^2}{2}\ell^2(d_2 + 3)^3 \\ & \stackrel{(ii)}{\leq} (d_1 + d_2)\ell^2\delta^2 + \frac{8\sigma^2}{S_1} + \frac{\mu_1^2}{2}\ell^2(d_1 + 3)^3 + \frac{\mu_2^2}{2}\ell^2(d_2 + 3)^3, \end{aligned}$$

where (i) follows from Lemma 14, and (ii) follows from eq. (45) and eq. (46).  $\square$

We denote

$$G_{\mu_1}(x, y, \nu_i, \xi_i) = \frac{F(x + \mu_1\nu_i, y, \xi_i) - F(x, y, \xi_i)}{\mu_1}\nu_i$$

and

$$H_{\mu_2}(x, y, \omega_i, \xi_i) = \frac{F(x, y + \mu_2\omega_i, \xi_i) - F(x, y, \xi_i)}{\mu_2}\omega_i$$

as unbiased estimators of  $\nabla_x f_{\mu_1}(x, y)$  and  $\nabla_y f_{\mu_2}(x, y)$ , respectively. Then we have the following lemma.

**Lemma 16.** *Suppose Assumption 2 holds, and suppose  $u_1$  and  $u_2$  are standard Gaussian random vector, i.e.,  $\nu_i \sim N(0, \mathbf{1}_{d_1})$  and  $\omega_i \sim N(0, \mathbf{1}_{d_2})$ . Then, we have*

$$\mathbb{E}[\|G_{\mu_1}(x, y, \nu_i, \xi_i) - G_{\mu_1}(x', y, \nu_i, \xi_i)\|_2^2] \leq 2(d_1 + 4)\ell^2\|x - x'\|_2^2 + 2\mu_1^2(d_1 + 6)^3\ell^2,$$

$$\mathbb{E} \left[ \|G_{\mu_1}(x, y, \nu_i, \xi_i) - G_{\mu_1}(x, y', \nu_i, \xi_i)\|_2^2 \right] \leq 2(d_1 + 4)\ell^2 \|y - y'\|_2^2 + 2\mu_1^2(d_1 + 6)^3\ell^2,$$

and

$$\mathbb{E} \left[ \|H_{\mu_2}(x, y, \nu_i, \xi_i) - H_{\mu_2}(x', y, \nu_i, \xi_i)\|_2^2 \right] \leq 2(d_2 + 4)\ell^2 \|x - x'\|_2^2 + 2\mu_2^2(d_2 + 6)^3\ell^2,$$

$$\mathbb{E} \left[ \|H_{\mu_2}(x, y, \nu_i, \xi_i) - H_{\mu_2}(x, y', \nu_i, \xi_i)\|_2^2 \right] \leq 2(d_2 + 4)\ell^2 \|y - y'\|_2^2 + 2\mu_2^2(d_2 + 6)^3\ell^2.$$

*Proof.* The proof is similar to that of Lemma 3 in Fang et al. (2018). Here we provide the proof for completeness. We only show how to upper bound the term  $\mathbb{E} \left[ \|G_{\mu_1}(x, y, \nu_1, \xi) - G_{\mu_1}(x', y, \nu_1, \xi)\|_2^2 \right]$  here. Then, the upper bounds on the remaining three terms can be obtained by following similar steps. We have that

$$\begin{aligned} & \mathbb{E} \left[ \|G_{\mu_1}(x, y, \nu_i, \xi_i) - G_{\mu_1}(x, y', \nu_i, \xi_i)\|_2^2 \right] \\ &= \mathbb{E} \left[ \left\| \frac{F(x + \mu_1 \nu_i, y, \xi_i) - F(x, y, \xi_i)}{\mu_1} \nu_1 - \frac{F(x + \mu_1 \nu_i, y', \xi_i) - F(x, y', \xi_i)}{\mu_1} \nu_1 \right\|_2^2 \right] \\ &= \mathbb{E} \left[ \left\| \frac{F(x + \mu_1 \nu_i, y, \xi_i) - F(x, y, \xi_i) - \langle \nabla_x F(x, y, \xi_i), \mu_1 \nu_i \rangle}{\mu_1} \nu_i \right. \right. \\ &\quad \left. \left. - \frac{F(x + \mu_1 \nu_i, y', \xi_i) - F(x, y', \xi_i) - \langle \nabla_x F(x, y', \xi_i), \mu_1 \nu_i \rangle}{\mu_1} \nu_i \right. \right. \\ &\quad \left. \left. + \langle \nabla_x F(x, y, \xi_i) - \nabla_x F(x, y', \xi_i), \nu_i \rangle \nu_i \right\|_2^2 \right] \\ &\leq 2\mathbb{E} \left[ \left\| \frac{F(x + \mu_1 \nu_i, y, \xi_i) - F(x, y, \xi_i) - \langle \nabla_x F(x, y, \xi_i), \mu_1 \nu_i \rangle}{\mu_1} \nu_i \right. \right. \\ &\quad \left. \left. - \frac{F(x + \mu_1 \nu_i, y', \xi_i) - F(x, y', \xi_i) - \langle \nabla_x F(x, y', \xi_i), \mu_1 \nu_i \rangle}{\mu_1} \nu_i \right\|_2^2 \right] \\ &\quad + 2\mathbb{E} \left[ \|\langle \nabla_x F(x, y, \xi_i) - \nabla_x F(x, y', \xi_i), \nu_i \rangle \nu_i\|_2^2 \right] \\ &\leq 4\mathbb{E} \left[ \left\| \frac{F(x + \mu_1 \nu_i, y, \xi_i) - F(x, y, \xi_i) - \langle \nabla_x F(x, y, \xi_i), \mu_1 \nu_i \rangle}{\mu_1} \nu_i \right\|_2^2 \right] \\ &\quad + 4\mathbb{E} \left[ \left\| \frac{F(x + \mu_1 \nu_i, y', \xi_i) - F(x, y', \xi_i) - \langle \nabla_x F(x, y', \xi_i), \mu_1 \nu_i \rangle}{\mu_1} \nu_i \right\|_2^2 \right] \\ &\quad + 2\mathbb{E} \left[ \|\langle \nabla_x F(x, y, \xi_i) - \nabla_x F(x, y', \xi_i), \nu_i \rangle \nu_i\|_2^2 \right] \\ &\leq 4\mathbb{E} \left[ \left| \frac{F(x + \mu_1 \nu_i, y, \xi_i) - F(x, y, \xi_i) - \langle \nabla_x F(x, y, \xi_i), \mu_1 \nu_i \rangle}{\mu_1} \right|^2 \|\nu_i\|_2^2 \right] \\ &\quad + 4\mathbb{E} \left[ \left| \frac{F(x + \mu_1 \nu_i, y', \xi_i) - F(x, y', \xi_i) - \langle \nabla_x F(x, y', \xi_i), \mu_1 \nu_i \rangle}{\mu_1} \right|^2 \|\nu_i\|_2^2 \right] \\ &\quad + 2\mathbb{E} \left[ \|\langle \nabla_x F(x, y, \xi_i) - \nabla_x F(x, y', \xi_i), \nu_i \rangle \nu_i\|_2^2 \right] \\ &\stackrel{(i)}{\leq} 2\mu_1^2 \ell^2 \mathbb{E} [\|\nu_i\|_2^2] + 2\mathbb{E} [\|\langle \nabla_x F(x, y, \xi_i) - \nabla_x F(x, y', \xi_i), \nu_i \rangle \nu_i\|_2^2] \\ &\stackrel{(ii)}{\leq} 2\mu_1^2 \ell^2 \mathbb{E} [\|\nu_i\|_2^2] + 2(d_1 + 4)\mathbb{E} [\|\nabla_x F(x, y, \xi_i) - \nabla_x F(x, y', \xi_i)\|_2^2] \\ &\stackrel{(iii)}{\leq} 2\mu_1^2 (d_1 + 6)^3 \ell^2 + 2(d_1 + 4)\ell^2 \mathbb{E} [\|y - y'\|_2^2], \end{aligned}$$

where (i) follows from the fact that for any  $a, a' \in \mathbb{R}^{d_1}$  and  $b \in \mathbb{R}^{d_2}$ , we have

$$|F(a, b, \xi_i) - F(a', b, \xi_i) - \langle \nabla_x F(a, b, \xi_i), a - a' \rangle| \leq \frac{\ell}{2} \|a - a'\|_2^2,$$

because  $F(a, b, \xi)$  has  $\ell$ -Lipschitz continuous gradient; (ii) follows because

$$\mathbb{E}[\|\langle a, \nu_i \rangle \nu_i\|_2^2] \leq (d_1 + 4) \|a\|_2^2,$$

obtained from (33) in [Nesterov & Spokoiny \(2017\)](#), and (iii) follows because  $\mathbb{E}[\|\nu_i\|_2^2] \leq (d_1 + 6)^3$  in (17) of [Nesterov & Spokoiny \(2017\)](#).  $\square$

### C.3 USEFUL PROPERTIES FOR ZERO-TH ORDER CONCAVE MAXIMIZER

In this section, we show some properties for the zeroth-order concave maximizer [Algorithm 5](#). For simplicity, for any given  $t \geq 0$ , we define  $g_t(y) = -f(x_{t+1}, y)$  and  $g_{t,\mu_2}(y) = -f_{\mu_2}(x_{t+1}, y)$ . [Lemma 11](#) and [Lemma 12](#) imply that  $g_t(\cdot)$  is  $\mu$ -strongly convex and has  $\ell$ -Lipschitz gradient, and  $g_{t,\mu_2}(\cdot)$  is convex and has  $\ell$ -Lipschitz gradient. We also define  $\tilde{y}_t^* = \operatorname{argmin}_y g_t(y)$ . We can obtain the following lemmas by following the same steps in [Luo et al. \(2020\)](#)

**Lemma 17** (Lemma 9 of [Luo et al. \(2020\)](#)). *Consider [Algorithm 5](#). We have*

$$\sum_{k=0}^m \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,k})\|_2^2] \leq \frac{2}{\beta} \mathbb{E}[g_{t,\mu_2}(\tilde{y}_{t,0}) - g_{t,\mu_2}(\tilde{y}_{t,m+1})] + \sum_{k=0}^m \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,k}) - \tilde{u}_{t,k}\|_2^2].$$

**Lemma 18** (Lemma 11 of [Luo et al. \(2020\)](#)). *Consider [Algorithm 5](#) with any  $\beta \leq \frac{2}{\ell}$  and  $k \geq 1$ . We have*

$$\mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,k}) - \tilde{u}_{t,k}\|_2^2] \leq \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,0}) - \tilde{u}_{t,0}\|_2^2] + \frac{\ell\beta}{2 - \ell\beta} \mathbb{E}[\|\tilde{u}_{t,0}\|_2^2].$$

**Lemma 19.** *Consider [Algorithm 5](#). For any  $k \geq 1$  and  $\beta \leq \frac{1}{\ell}$ , we have*

$$\begin{aligned} \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,\tilde{m}_t})\|_2^2] &\leq \frac{2}{\beta\mu(m+1)} \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,0})\|_2^2] + \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,0}) - \tilde{u}_{t,0}\|_2^2] + \frac{\ell\beta}{2 - \ell\beta} \mathbb{E}[\|\tilde{u}_{t,0}\|_2^2] \\ &\quad + \frac{2}{\beta(m+1)} \left( \frac{\mu_2^2}{4\mu} \ell^2 (d_2 + 3)^3 + \mu_2^2 \ell d_2 \right). \end{aligned}$$

*Proof.* Taking summation of the result of [Lemma 18](#) over  $t = \{0, \dots, m\}$  yields

$$\sum_{k=0}^m \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,k}) - \tilde{u}_{t,k}\|_2^2] \leq (m+1) \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,0}) - \tilde{u}_{t,0}\|_2^2] + \frac{\ell\beta(m+1)}{2 - \ell\beta} \mathbb{E}[\|\tilde{u}_{t,0}\|_2^2]. \quad (47)$$

Combining eq. (47) with [Lemma 17](#) yields

$$\begin{aligned} \sum_{k=0}^m \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,k})\|_2^2] &\leq \frac{2}{\beta} \mathbb{E}[g_{t,\mu_2}(\tilde{y}_{t,0}) - g_{t,\mu_2}(\tilde{y}_{t,m+1})] + (m+1) \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,0}) - \tilde{u}_{t,0}\|_2^2] \\ &\quad + \frac{\ell\beta(m+1)}{2 - \ell\beta} \mathbb{E}[\|\tilde{u}_{t,0}\|_2^2]. \end{aligned} \quad (48)$$

Dividing both sides of eq. (48) yields

$$\begin{aligned} \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,\tilde{m}_t})\|_2^2] &\leq \frac{2}{\beta(m+1)} \mathbb{E}[g_{t,\mu_2}(\tilde{y}_{t,0}) - g_{t,\mu_2}(\tilde{y}_{t,m+1})] + \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,0}) - \tilde{u}_{t,0}\|_2^2] \\ &\quad + \frac{\ell\beta}{2 - \ell\beta} \mathbb{E}[\|\tilde{u}_{t,0}\|_2^2]. \end{aligned} \quad (49)$$

We can bound the term  $\mathbb{E}[g_{t,\mu_2}(\tilde{y}_{t,0}) - g_{t,\mu_2}(\tilde{y}_{t,m+1})]$  as follows:

$$\mathbb{E}[g_{t,\mu_2}(\tilde{y}_{t,0}) - g_{t,\mu_2}(\tilde{y}_{t,m+1})]$$

$$\begin{aligned}
&= \mathbb{E}[g_t(\tilde{y}_{t,0}) - g_t(\tilde{y}_{t,m+1})] + \mathbb{E}[g_{t,\mu_2}(\tilde{y}_{t,0}) - g_t(\tilde{y}_{t,0})] + \mathbb{E}[g_t(\tilde{y}_{t,m+1}) - g_{t,\mu_2}(\tilde{y}_{t,m+1})] \\
&\leq \mathbb{E}[g_t(\tilde{y}_{t,0}) - g_t(\tilde{y}_{t,m+1})] + \mathbb{E}[|g_{t,\mu_2}(\tilde{y}_{t,0}) - g_t(\tilde{y}_{t,0})|] + \mathbb{E}[|g_{t,\mu_2}(\tilde{y}_{t,m+1}) - g_t(\tilde{y}_{t,m+1})|] \\
&\stackrel{(i)}{\leq} \mathbb{E}[g_t(\tilde{y}_{t,0}) - g_t(\tilde{y}_{t,m+1})] + \mu_2^2 \ell d_2 \\
&\leq \mathbb{E}[g_t(\tilde{y}_{t,0}) - g_t(\tilde{y}_t^*)] + \mu_2^2 \ell d_2 \\
&\stackrel{(ii)}{\leq} \frac{1}{2\mu} \mathbb{E}[\|\nabla g_t(\tilde{y}_{t,0})\|_2^2] + \mu_2^2 \ell d_2 \\
&\leq \frac{1}{\mu} \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,0})\|_2^2] + \frac{1}{\mu} \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,0}) - \nabla g_t(\tilde{y}_{t,0})\|_2^2] + \mu_2^2 \ell d_2 \\
&\stackrel{(iii)}{\leq} \frac{1}{\mu} \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,0})\|_2^2] + \frac{\mu_2^2}{4\mu} \ell^2 (d_2 + 3)^3 + \mu_2^2 \ell d_2, \tag{50}
\end{aligned}$$

where (i) follows from Lemma 13, (ii) follows from eq. (6) in Lemma 2, and (iii) follows from Lemma 14. Substituting eq. (50) into eq. (49) yields

$$\begin{aligned}
\mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,\tilde{m}_t})\|_2^2] &\leq \frac{2}{\beta\mu(m+1)} \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,0})\|_2^2] + \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_{t,0}) - \tilde{u}_{t,0}\|_2^2] + \frac{\ell\beta}{2-\ell\beta} \mathbb{E}[\|\tilde{u}_{t,0}\|_2^2] \\
&\quad + \frac{2}{\beta(m+1)} \left( \frac{\mu_2^2}{4\mu} \ell^2 (d_2 + 3)^3 + \mu_2^2 \ell d_2 \right),
\end{aligned}$$

which completes the proof.  $\square$

**Lemma 20.** Consider Algorithm 5. Let  $S_{2,y} \geq 16\kappa(d_2 + 4)\ell\beta$  and  $\beta \leq \frac{1}{6\ell}$ . For any  $t > 0$ , we have

$$\sum_{k=0}^m \mathbb{E}[\|\tilde{u}_{t,k}\|_2^2] \leq \frac{1}{1-b} \mathbb{E}[\|\tilde{u}_{t,0}\|_2^2] + \frac{m+1}{1-b} \left[ \frac{2\mu_2^2 \ell \kappa}{\beta} (d_2 + 3)^3 + 7\mu_2^2 (d_2 + 6)^3 \ell^2 \right],$$

where  $b = 1 - \frac{\beta\mu\ell}{2(\mu+\ell)}$ .

*Proof.* The update of Algorithm 5 implies that

$$\begin{aligned}
&\mathbb{E}[\|\tilde{u}_{t,k}\|_2^2 | \mathcal{F}_{t,k}] \\
&= \|\tilde{u}_{t,k-1}\|_2^2 + 2\mathbb{E}[\langle \tilde{u}_{t,k-1}, H_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}, \omega_{\mathcal{M}_2}, \xi_{\mathcal{M}}) - H_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}, \omega_{\mathcal{M}_2}, \xi_{\mathcal{M}}) \rangle | \mathcal{F}_{t,k}] \\
&\quad + \mathbb{E}[\|H_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}, \omega_{\mathcal{M}_2}, \xi_{\mathcal{M}}) - H_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}, \omega_{\mathcal{M}_2}, \xi_{\mathcal{M}})\|_2^2 | \mathcal{F}_{t,k}] \\
&= \|\tilde{u}_{t,k-1}\|_2^2 + \frac{2}{\beta} \langle \tilde{y}_{t,k} - \tilde{y}_{t,k-1}, \nabla_y f_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \nabla_y f_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}) \rangle \\
&\quad + \mathbb{E}[\|H_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}, \omega_{\mathcal{M}_2}, \xi_{\mathcal{M}}) - H_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}, \omega_{\mathcal{M}_2}, \xi_{\mathcal{M}})\|_2^2 | \mathcal{F}_{t,k}] \\
&= \|\tilde{u}_{t,k-1}\|_2^2 + \frac{2}{\beta} \langle \tilde{y}_{t,k} - \tilde{y}_{t,k-1}, \nabla_y f(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \nabla_y f(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}) \rangle \\
&\quad + \frac{2}{\beta} \langle \tilde{y}_{t,k} - \tilde{y}_{t,k-1}, \nabla_y f_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \nabla_y f(\tilde{x}_{t,k}, \tilde{y}_{t,k}) \rangle \\
&\quad + \frac{2}{\beta} \langle \tilde{y}_{t,k} - \tilde{y}_{t,k-1}, \nabla_y f(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}) - \nabla_y f_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}) \rangle \\
&\quad + \mathbb{E}[\|H_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}, \omega_{\mathcal{M}_2}, \xi_{\mathcal{M}}) - H_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}, \omega_{\mathcal{M}_2}, \xi_{\mathcal{M}})\|_2^2 | \mathcal{F}_{t,k}] \\
&\stackrel{(i)}{\leq} \|\tilde{u}_{t,k-1}\|_2^2 - \frac{2}{\beta} \left( \frac{\mu\ell}{\mu+\ell} \|\tilde{y}_{t,k} - \tilde{y}_{t,k-1}\|_2^2 + \frac{1}{\mu+\ell} \|\nabla_y f(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \nabla_y f(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1})\|_2^2 \right) \\
&\quad + \frac{2}{\beta} \left( \frac{\mu\ell}{4(\mu+\ell)} \|\tilde{y}_{t,k} - \tilde{y}_{t,k-1}\|_2^2 + \frac{\mu+\ell}{\mu\ell} \|\nabla_y f_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \nabla_y f(\tilde{x}_{t,k}, \tilde{y}_{t,k})\|_2^2 \right) \\
&\quad + \frac{2}{\beta} \left( \frac{\mu\ell}{4(\mu+\ell)} \|\tilde{y}_{t,k} - \tilde{y}_{t,k-1}\|_2^2 + \frac{\mu+\ell}{\mu\ell} \|\nabla_y f_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}) - \nabla_y f(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1})\|_2^2 \right) \\
&\quad + \mathbb{E}[\|H_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}, \omega_{\mathcal{M}_2}, \xi_{\mathcal{M}}) - H_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}, \omega_{\mathcal{M}_2}, \xi_{\mathcal{M}})\|_2^2 | \mathcal{F}_{t,k}]
\end{aligned}$$

$$\begin{aligned}
& \stackrel{(ii)}{\leq} \|\tilde{u}_{t,k-1}\|_2^2 - \frac{\mu\ell}{\beta(\mu+\ell)} \|\tilde{y}_{t,k} - \tilde{y}_{t,k-1}\|_2^2 - \frac{2}{\beta(\mu+\ell)} \|\nabla_y f(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \nabla_y f(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1})\|_2^2 \\
& \quad + \frac{\mu_2^2 \ell (\mu + \ell)}{\beta \mu} (d_2 + 3)^3 + \mathbb{E}[\|H_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}, \omega_{\mathcal{M}_2}, \xi_{\mathcal{M}}) - H_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}, \omega_{\mathcal{M}_2}, \xi_{\mathcal{M}})\|_2^2 | \mathcal{F}_{t,k}] \\
& \leq \left(1 - \frac{\beta\mu\ell}{\mu + \ell}\right) \|\tilde{u}_{t,k-1}\|_2^2 - \frac{2}{\beta(\mu + \ell)} \|\nabla_y f(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \nabla_y f(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1})\|_2^2 \\
& \quad + 2\mathbb{E}[\|H_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}, \omega_{\mathcal{M}_2}, \xi_{\mathcal{M}}) - H_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}, \omega_{\mathcal{M}_2}, \xi_{\mathcal{M}}) \\
& \quad \quad - (\nabla_y f_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \nabla_y f_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}))\|_2^2 | \mathcal{F}_{t,k}] \\
& \quad + 2\mathbb{E}[\|\nabla_y f_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \nabla_y f_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1})\|_2^2 | \mathcal{F}_{t,k}] + \frac{\mu_2^2 \ell (\mu + \ell)}{\beta \mu} (d_2 + 3)^3 \\
& \leq \left(1 - \frac{\beta\mu\ell}{\mu + \ell}\right) \|\tilde{u}_{t,k-1}\|_2^2 - \frac{2}{\beta(\mu + \ell)} \|\nabla_y f(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \nabla_y f(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1})\|_2^2 \\
& \quad + 2\mathbb{E}[\|H_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}, \omega_{\mathcal{M}_2}, \xi_{\mathcal{M}}) - H_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}, \omega_{\mathcal{M}_2}, \xi_{\mathcal{M}}) \\
& \quad \quad - (\nabla_y f_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \nabla_y f_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}))\|_2^2 | \mathcal{F}_{t,k}] \\
& \quad + 6\mathbb{E}[\|\nabla_y f(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \nabla_y f(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1})\|_2^2 | \mathcal{F}_{t,k}] \\
& \quad + 6\mathbb{E}[\|\nabla_y f_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}) - \nabla_y f_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1})\|_2^2 | \mathcal{F}_{t,k}] \\
& \quad + 6\mathbb{E}[\|\nabla_y f_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \nabla_y f_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k})\|_2^2 | \mathcal{F}_{t,k}] + \frac{\mu_2^2 \ell (\mu + \ell)}{\beta \mu} (d_2 + 3)^3 \\
& \leq \left(1 - \frac{\beta\mu\ell}{\mu + \ell}\right) \|\tilde{u}_{t,k-1}\|_2^2 - \left(\frac{2}{\beta(\mu + \ell)} - 6\right) \|\nabla_y f(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \nabla_y f(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1})\|_2^2 \\
& \quad + \frac{2}{S_{2,y}} \mathbb{E}[\|H_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}, \omega_i, \xi_i) - H_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}, \omega_i, \xi_i)\|_2^2 | \mathcal{F}_{t,k}] \\
& \quad + 3\mu_2^2 \ell^2 (d_2 + 3)^3 + \frac{\mu_2^2 \ell (\mu + \ell)}{\beta \mu} (d_2 + 3)^3 \\
& \stackrel{(iv)}{\leq} \left(1 - \frac{\beta\mu\ell}{\mu + \ell}\right) \|\tilde{u}_{t,k-1}\|_2^2 + \frac{2}{S_{2,y}} \mathbb{E}[\|H_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}, \omega_i, \xi_i) - H_{\mu_2}(\tilde{x}_{t,k-1}, \tilde{y}_{t,k-1}, \omega_i, \xi_i)\|_2^2 | \mathcal{F}_{t,k}] \\
& \quad + 3\mu_2^2 \ell^2 (d_2 + 3)^3 + \frac{\mu_2^2 \ell (\mu + \ell)}{\beta \mu} (d_2 + 3)^3 \\
& \stackrel{(v)}{\leq} \left(1 - \frac{\beta\mu\ell}{\mu + \ell}\right) \|\tilde{u}_{t,k-1}\|_2^2 + \frac{2}{S_{2,y}} \left[2(d_2 + 4)\ell^2 \beta^2 \|\tilde{u}_{t,k-1}\|_2^2 + 2\mu_2^2 (d_2 + 6)^3 \ell^2\right] \\
& \quad + 3\mu_2^2 \ell^2 (d_2 + 3)^3 + \frac{\mu_2^2 \ell (\mu + \ell)}{\beta \mu} (d_2 + 3)^3 \\
& = \left(1 - \frac{\beta\mu\ell}{\mu + \ell} + \frac{4}{S_{2,y}} (d_2 + 4)\ell^2 \beta^2\right) \|\tilde{u}_{t,k-1}\|_2^2 \\
& \quad + \frac{4}{S_{2,y}} \mu_2^2 (d_2 + 6)^3 \ell^2 + 3\mu_2^2 \ell^2 (d_2 + 3)^3 + \frac{\mu_2^2 \ell (\mu + \ell)}{\beta \mu} (d_2 + 3)^3 \\
& \stackrel{(vi)}{\leq} \left(1 - \frac{\beta\mu\ell}{2(\mu + \ell)}\right) \|\tilde{u}_{t,k-1}\|_2^2 + \frac{\mu_2^2 \ell (1 + \kappa)}{\beta} (d_2 + 3)^3 + 7\mu_2^2 (d_2 + 6)^3 \ell^2. \tag{51}
\end{aligned}$$

where (i) follows from eq. (4) in Lemma 2 and Yong's inequality, (ii) follows from Lemma 14, (iii) follows from Lemma 1 in Fang et al. (2018), (iv) follows from the fact that  $\frac{2}{\beta(\mu+\ell)} - 6 > 0$ , (v) follows from Lemma 16, and (vi) follows from the fact that  $\frac{4}{S_{2,y}} (d_2 + 4)\ell^2 \beta^2 \leq \frac{\beta\mu\ell}{2(\mu+\ell)}$ . Taking expectation on both sides of eq. (51) and applying eq. (51) iteratively yield

$$\mathbb{E}[\|\tilde{u}_{t,k}\|_2^2] \leq b^k \mathbb{E}[\|\tilde{u}_{t,0}\|_2^2] + \left[ \frac{2\mu_2^2 \ell \kappa}{\beta} (d_2 + 3)^3 + 7\mu_2^2 (d_2 + 6)^3 \ell^2 \right] \sum_{j=0}^{k-1} b^j. \tag{52}$$



Taking summation of eq. (52) over  $k = \{0, \dots, m\}$  yields

$$\begin{aligned} \sum_{k=0}^m \mathbb{E}[\|\tilde{u}_{t,k}\|_2^2] &\leq \mathbb{E}[\|\tilde{u}_{t,0}\|_2^2] \sum_{k=0}^m b^k + \left[ \frac{2\mu_2^2 \ell \kappa}{\beta} (d_2 + 3)^3 + 7\mu_2^2 (d_2 + 6)^3 \ell^2 \right] \sum_{k=0}^m \sum_{j=0}^{k-1} b^j \\ &\leq \frac{1}{1-b} \mathbb{E}[\|\tilde{u}_{t,0}\|_2^2] + \frac{m+1}{1-b} \left[ \frac{2\mu_2^2 \ell \kappa}{\beta} (d_2 + 3)^3 + 7\mu_2^2 (d_2 + 6)^3 \ell^2 \right], \end{aligned}$$

which completes the proof.  $\square$

#### C.4 INITIALIZATION BY ZERO-ORDER ISARAH

We present the detailed procedure of ZO-iSARAH in Algorithm 6, which is used to initialize  $y_0$  in ZO-SREDA-Boost (line 2 of Algorithm 4). We consider the following convex optimization problem:

$$\min_{w \in \mathbb{R}^d} p(w) \triangleq \mathbb{E}[P(w; \xi)], \quad (53)$$

where  $P$  is average  $\ell$ -gradient Lipschitz and convex,  $p$  is  $\mu$ -strongly convex, and  $\xi$  is a random vector. We define

$$\Psi_\tau(w, \psi_{\mathcal{M}_1}, \xi_{\mathcal{M}}) = \frac{1}{|\mathcal{M}|} \sum_{i \in [\mathcal{M}]} \frac{P(w + \tau \psi_i, \xi_i) - P(w, \xi_i)}{\tau} \psi_i, \quad (54)$$

where  $\psi_i \sim N(0, \mathbf{I}_d)$  independently across the index  $i$ .

---

#### Algorithm 6 ZO-iSARAH

---

```

1: Input:  $\tilde{w}_0$ , learning rate  $\gamma > 0$ , inner loop size  $I$ , batch size  $B_1$  and  $B_2$ 
2: for  $t = 1, 2, \dots, T$  do
3:    $w_0 = \tilde{w}_{t-1}$ 
4:   draw  $B_1$  samples  $\{\xi_1, \dots, \xi_{B_1}\}$ 
5:    $v_0 = \frac{1}{B_1} \sum_{i=1}^{B_1} \sum_{j=1}^d \frac{P(w_0 + \delta e_j, \xi_i) - P(w_0 - \delta e_j, \xi_i)}{2\delta} e_j$ 
6:   where  $e_j$  denotes the vector with  $j$ -th natural unit basis vector.
7:    $w_1 = w_0 + \gamma v_0$ 
8:   for  $k = 1, 2, \dots, I - 1$  do
9:     Draw minibatch sample  $\mathcal{M} = \{\xi_1, \dots, \xi_{B_2}\}$  and  $\mathcal{M}_1 = \{\psi_1, \dots, \psi_{B_2}\}$ 
10:     $v_k = v_{k-1} + \Psi_\tau(w_k, \psi_{\mathcal{M}_1}, \xi_{\mathcal{M}}) - \Psi_\tau(w_{k-1}, \psi_{\mathcal{M}_1}, \xi_{\mathcal{M}})$ 
11:     $w_{k+1} = w_k - \gamma v_k$ 
12:   end for
13:    $\tilde{w}_t$  chosen uniformly at random from  $\{w_k\}_{k=0}^I$ 
14: end for

```

---

We have the following convergence result by using ZO-iSARAH to solve problem eq. (53).

**Lemma 21.** Consider Algorithm 6. Set  $\gamma = \frac{2}{9\ell}$ ,  $B_1 = \frac{25\sigma^2}{\epsilon}$ ,  $B_2 = d$ ,  $I = 36\kappa - 1$ ,  $T = \log_2 \frac{5\|\nabla p_\tau(\tilde{w}_0)\|_2^2}{\epsilon}$ ,  $\delta = \frac{2\epsilon^{0.5}}{5\ell d^{0.6}}$ , and  $\tau = \min\{\frac{\epsilon^{0.5}}{3\ell(d+3)^{1.5}}, \sqrt{\frac{2\epsilon}{5\ell\mu d}}\}$ . Then, we have

$$\mathbb{E}[\|\nabla p_\tau(\tilde{w}_T)\|_2^2] \leq \epsilon,$$

with the total sample complexity given by  $\mathcal{O}\left((\kappa + \frac{1}{\epsilon}) \log\left(\frac{1}{\epsilon}\right)\right)$ .

*Proof.* Following steps similar to those in Lemmas 17-19, at  $t$ -th outer-loop iteration, we can obtain the following convergence result of inner loop:

$$\begin{aligned} &\mathbb{E}[\|\nabla p_\tau(\tilde{w}_t)\|_2^2] \\ &\leq \frac{2}{\gamma\tau(I+1)} \mathbb{E}[\|\nabla p_\tau(w_0)\|_2^2] + \mathbb{E}[\|\nabla p_\tau(w_0) - v_0\|_2^2] + \frac{\ell\gamma}{2-\ell\gamma} \mathbb{E}[\|v_0\|_2^2] \\ &\quad + \frac{2}{\gamma(I+1)} \left( \frac{\tau^2}{4\mu} \ell^2 (d+3)^3 + \tau^2 \ell d \right) \end{aligned}$$

$$\begin{aligned} &\leq \left( \frac{2}{\gamma\mu(I+1)} + \frac{2\ell\gamma}{2-\ell\gamma} \right) \mathbb{E}[\|\nabla p_\tau(w_0)\|_2^2] + \left( 1 + \frac{2\ell\gamma}{2-\ell\gamma} \right) \mathbb{E}[\|\nabla p_\tau(w_0) - v_0\|_2^2] \\ &\quad + \frac{2}{\gamma(I+1)} \left( \frac{\tau^2}{4\mu} \ell^2(d+3)^3 + \tau^2 \ell d \right). \end{aligned} \quad (55)$$

Then, following steps similar to those in Lemma 15, we can obtain

$$\mathbb{E}[\|\nabla p_\tau(w_0) - v_0\|_2^2] \leq \frac{2\sigma^2}{B_1} + \frac{d\ell^2\delta^2}{2} + \frac{\tau^2}{2} \ell^2(d+3)^3. \quad (56)$$

Letting  $\gamma = \frac{2}{9\ell}$ ,  $I = 36\kappa - 1$ , substituting eq. (56) into eq. (55), and recalling the fact that  $w_I = \tilde{w}_t$  and  $w_0 = \tilde{w}_{t-1}$  yield

$$\mathbb{E}[\|\nabla p_\tau(\tilde{w}_t)\|_2^2] \leq \frac{1}{2} \mathbb{E}[\|\nabla p_\tau(\tilde{w}_{t-1})\|_2^2] + \frac{5\sigma^2}{2B_1} + \frac{5d\ell^2\delta^2}{8} + \frac{11\tau^2}{16} \ell^2(d+3)^3 + \frac{\tau^2}{4} \ell\mu d. \quad (57)$$

Applying eq. (57) iteratively from  $t = T$  to 0 yields

$$\begin{aligned} \mathbb{E}[\|\nabla p_\tau(\tilde{w}_T)\|_2^2] &\leq \frac{1}{2^T} \|\nabla p_\tau(\tilde{w}_0)\|_2^2 + \frac{5\sigma^2}{2B_1} \sum_{t=0}^{T-1} \frac{1}{2^t} \\ &\quad + \left( \frac{5d\ell^2\delta^2}{8} + \frac{11\tau^2}{16} \ell^2(d+3)^3 + \frac{\tau^2}{4} \ell\mu d \right) \sum_{t=0}^{T-1} \frac{1}{2^t} \\ &\leq \frac{1}{2^T} \|\nabla p_\tau(\tilde{w}_0)\|_2^2 + \frac{5\sigma^2}{B_1} + \frac{5d\ell^2\delta^2}{4} + \frac{11\tau^2}{8} \ell^2(d+3)^3 + \frac{\tau^2}{2} \ell\mu d. \end{aligned} \quad (58)$$

Letting  $T = \log_2 \frac{5\|\nabla p_\tau(\tilde{w}_0)\|_2^2}{\epsilon}$ ,  $B_1 = \frac{25\sigma^2}{\epsilon}$ ,  $\delta = \frac{2\epsilon^{0.5}}{5\ell d^{0.5}}$ , and  $\tau = \min\{\frac{\epsilon^{0.5}}{3\ell(d+3)^{1.5}}, \sqrt{\frac{2\epsilon}{5\ell\mu d}}\}$ , we have

$$\mathbb{E}[\|\nabla p_\tau(\tilde{w}_T)\|_2^2] \leq \epsilon.$$

The total sample complexity is given by

$$T \cdot (I \cdot B_2 + d \cdot B_1) = \mathcal{O} \left( d \left( \kappa + \frac{1}{\epsilon} \right) \log \left( \frac{1}{\epsilon} \right) \right).$$

**Extension to finite-sum case:** ZO-iSARAH is also applicable for strongly-convex optimization in the finite-sum case, which takes the form given by

$$\min_{w \in \mathbb{R}^d} p(w) \triangleq \frac{1}{n} \sum_{i=1}^n P(w; \xi_i). \quad (59)$$

To solve the problem in eq. (59), we slightly modify Algorithm 6 by replacing line 5 with the full gradient. Following steps similar to those from eq. (55) to eq. (58), we have

$$\mathbb{E}[\|\nabla p_\tau(\tilde{w}_T)\|_2^2] \leq \frac{1}{2^T} \|\nabla p_\tau(\tilde{w}_0)\|_2^2 + \frac{5d\ell^2\delta^2}{4} + \frac{11\tau^2}{8} \ell^2(d+3)^3 + \frac{\tau^2}{2} \ell\mu d.$$

Letting  $T = \log_2 \frac{4\|\nabla p_\tau(\tilde{w}_0)\|_2^2}{\epsilon}$ ,  $\delta = \frac{\epsilon^{0.5}}{3\ell d^{0.5}}$ , and  $\tau = \min\{\frac{\epsilon^{0.5}}{3\ell(d+3)^{1.5}}, \sqrt{\frac{\epsilon}{2\ell\mu d}}\}$ , we have

$$\mathbb{E}[\|\nabla p_\tau(\tilde{w}_T)\|_2^2] \leq \epsilon.$$

The total sample complexity is given by

$$T \cdot (I \cdot B_2 + d \cdot n) = \mathcal{O} \left( d (\kappa + n) \log \left( \frac{1}{\epsilon} \right) \right). \quad (60)$$

□

Let  $P(\cdot; \xi) = -F(x_0, \cdot; \xi)$ . Then we can conclude that the sample complexity for the initialization of Algorithm 4 is given by  $\mathcal{O}(d_2\kappa \log(\kappa))$  in the online case, and given by  $\mathcal{O}(d_2(\kappa + n) \log(\kappa))$  in the finite-sum case.

## C.5 PROOF OF THEOREM 2

We define  $\Delta'_t = \mathbb{E}[\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|_2^2] + \mathbb{E}[\|\nabla_y f_{\mu_2}(x_t, y_t) - u_t\|_2^2]$ ,  $\tilde{\Delta}'_{t,k} = \mathbb{E}[\|\nabla_x f_{\mu_1}(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \tilde{v}_{t,k}\|_2^2] + \mathbb{E}[\|\nabla_y f_{\mu_2}(\tilde{x}_{t,k}, \tilde{y}_{t,k}) - \tilde{u}_{t,k}\|_2^2]$ , and  $\delta'_t = \mathbb{E}[\|\nabla_y f_{\mu_2}(x_t, y_t)\|_2^2]$ . In this section, we establish the following lemmas to characterize the relationship between  $\Delta_t$  and  $\Delta'_t$ , and  $\delta_t$  and  $\delta'_t$ , and the recursive relationship of  $\Delta'_t$  and  $\delta'_t$ , which are crucial for the analysis of Theorem 2.

**Lemma 22.** *Suppose Assumption 2 hold. Then, for any  $0 \leq t \leq T - 1$ , we have*

$$\Delta_t \leq 2\Delta'_t + \frac{\mu_1^2}{2}\ell^2(d_1 + 3)^3 + \frac{\mu_2^2}{2}\ell^2(d_2 + 3)^3,$$

and

$$\delta_t \leq 2\delta'_t + \frac{\mu_2^2}{2}\ell^2(d_2 + 3)^3.$$

*Proof.* For the first inequality, we have

$$\begin{aligned} \Delta_t &= \mathbb{E}[\|\nabla_x f(x_t, y_t) - v_t\|_2^2] + \mathbb{E}[\|\nabla_y f(x_t, y_t) - u_t\|_2^2] \\ &= \mathbb{E}[\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t + \nabla_x f(x_t, y_t) - \nabla_x f_{\mu_1}(x_t, y_t)\|_2^2] \\ &\quad + \mathbb{E}[\|\nabla_y f_{\mu_2}(x_t, y_t) - u_t + \nabla_y f(x_t, y_t) - \nabla_y f_{\mu_2}(x_t, y_t)\|_2^2] \\ &\leq 2\mathbb{E}[\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|_2^2] + 2\mathbb{E}[\|\nabla_y f_{\mu_2}(x_t, y_t) - u_t\|_2^2] \\ &\quad + 2\mathbb{E}[\|\nabla_x f(x_t, y_t) - \nabla_x f_{\mu_1}(x_t, y_t)\|_2^2] + 2\mathbb{E}[\|\nabla_y f(x_t, y_t) - \nabla_y f_{\mu_2}(x_t, y_t)\|_2^2] \\ &\stackrel{(i)}{\leq} 2\Delta'_t + \frac{\mu_1^2}{2}\ell^2(d_1 + 3)^3 + \frac{\mu_2^2}{2}\ell^2(d_2 + 3)^3, \end{aligned}$$

where (i) follows from Lemma 14. For the second inequality, we have

$$\begin{aligned} \delta_t &= \mathbb{E}[\|\nabla_y f(x_t, y_t)\|_2^2] = \mathbb{E}[\|\nabla_y f_{\mu_2}(x_t, y_t) + \nabla_y f(x_t, y_t) - \nabla_y f_{\mu_2}(x_t, y_t)\|_2^2] \\ &\leq 2\mathbb{E}[\|\nabla_y f_{\mu_2}(x_t, y_t)\|_2^2] + 2\mathbb{E}[\|\nabla_y f(x_t, y_t) - \nabla_y f_{\mu_2}(x_t, y_t)\|_2^2] \\ &\stackrel{(i)}{\leq} 2\delta'_t + \frac{\mu_2^2}{2}\ell^2(d_2 + 3)^3, \end{aligned}$$

where (i) follows from Lemma 14. □

We provide the following two lemmas to characterize the relationship between  $\delta'_t$  and  $\delta'_{t-1}$  as well as that between  $\Delta'_t$  and  $\Delta'_{t-1}$ .

**Lemma 23.** *Suppose Assumption 2 hold. Then, we have*

$$\begin{aligned} \Delta'_t &\leq \left[1 + \frac{6\ell^2\beta^2}{1-b} \left(\frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}}\right)\right] \Delta'_{t-1} + \frac{6\ell^2\beta^2}{1-b} \left(\frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}}\right) \delta'_{t-1} \\ &\quad + 2\ell^2\alpha^2 \left(\frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}}\right) \left(1 + \frac{9\ell^2\beta^2}{1-b}\right) \mathbb{E}[\|v_{t-1}\|_2^2] + \pi_\Delta(d_1, d_2, \mu_1, \mu_2), \end{aligned}$$

where  $b = 1 - \frac{\beta\mu\ell}{2(\mu+\ell)}$  and

$$\begin{aligned} \pi_\Delta(d_1, d_2, \mu_1, \mu_2) &= \frac{2\ell^2\beta^2}{1-b} \left(\frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}}\right) \left\{ 6\ell^2 \left[ \frac{\mu_1^2(d_1+6)^3}{S_{2,x}} + \frac{\mu_2^2(d_2+6)^3}{S_{2,y}} \right] + (m+1) \left( \frac{2\mu_2^2\ell\kappa}{\beta} (d_2+3)^3 \right. \right. \\ &\quad \left. \left. + 7\mu_2^2(d_2+6)^3\ell^2 \right) \right\} + \frac{2(m+2)\mu_1^2(d_1+6)^3\ell^2}{S_{2,x}} + \frac{2(m+2)\mu_2^2(d_2+6)^3\ell^2}{S_{2,y}}. \end{aligned}$$

Moreover, if we let  $\beta = \frac{2}{13\ell}$ ,  $m = 104\kappa - 1$ ,  $S_{2,x} \geq 5600(d_1 + 4)$  and  $S_{2,y} \geq 5600(d_2 + 4)$ , then we have

$$\pi_\Delta(d_1, d_2, \mu_1, \mu_2) \leq \kappa^3\ell^2[\mu_1^2(d_1+6)^3 + \mu_2^2(d_2+6)^3].$$

*Proof.* We proceed as follows:

$$\begin{aligned}
\Delta'_t &= \tilde{\Delta}'_{t-1, \tilde{m}_{t-1}} \\
&= \mathbb{E} \left[ \left\| \nabla_x f_{\mu_1}(\tilde{x}_{t-1, \tilde{m}_{t-1}}, \tilde{y}_{t-1, \tilde{m}_{t-1}}) - \tilde{v}_{t-1, \tilde{m}_{t-1}} \right\|_2^2 \right] \\
&\stackrel{(i)}{\leq} \mathbb{E} \left[ \left\| \nabla_x f_{\mu_1}(\tilde{x}_{t-1, \tilde{m}_{t-1}-1}, \tilde{y}_{t-1, \tilde{m}_{t-1}-1}) - \tilde{v}_{t-1, \tilde{m}_{t-1}-1} \right\|_2^2 \right] \\
&\quad + \frac{1}{S_{2,x}} \mathbb{E} \left[ \left\| G_{\mu_1}(\tilde{x}_{t-1, \tilde{m}_{t-1}}, \tilde{y}_{t-1, \tilde{m}_{t-1}}, \nu_i, \xi_i) - G_{\mu_1}(\tilde{x}_{t-1, \tilde{m}_{t-1}-1}, \tilde{y}_{t-1, \tilde{m}_{t-1}-1}, \nu_i, \xi_i) \right\|_2^2 \right] \\
&\stackrel{(ii)}{\leq} \mathbb{E} \left[ \left\| \nabla_x f_{\mu_1}(\tilde{x}_{t-1, \tilde{m}_{t-1}-1}, \tilde{y}_{t-1, \tilde{m}_{t-1}-1}) - \tilde{v}_{t-1, \tilde{m}_{t-1}-1} \right\|_2^2 \right] \\
&\quad + \frac{1}{S_{2,x}} \left[ 2(d_1 + 4)\ell^2\beta^2 \mathbb{E}[\|\tilde{u}_{t-1, \tilde{m}_{t-1}-1}\|_2^2] + 2\mu_1^2(d_1 + 6)^3\ell^2 \right], \tag{61}
\end{aligned}$$

where (i) follows from Lemma 5, and (ii) follows from Lemma 16. Applying eq. (61) recursively yields

$$\begin{aligned}
&\mathbb{E} \left[ \left\| \nabla_x f_{\mu_1}(\tilde{x}_{t-1, \tilde{m}_{t-1}}, \tilde{y}_{t-1, \tilde{m}_{t-1}}) - \tilde{v}_{t-1, \tilde{m}_{t-1}} \right\|_2^2 \right] \\
&\leq \mathbb{E} \left[ \left\| \nabla_x f_{\mu_1}(\tilde{x}_{t-1,0}, \tilde{y}_{t-1,0}) - \tilde{v}_{t-1,0} \right\|_2^2 \right] + \frac{2(d_1 + 4)\ell^2\beta^2}{S_{2,x}} \sum_{k=0}^{\tilde{m}_{t-1}-1} \mathbb{E}[\|\tilde{u}_{t-1,k}\|_2^2] \\
&\quad + \frac{2\tilde{m}_{t-1}\mu_1^2(d_1 + 6)^3\ell^2}{S_{2,x}} \\
&\leq \mathbb{E} \left[ \left\| \nabla_x f_{\mu_1}(\tilde{x}_{t-1,0}, \tilde{y}_{t-1,0}) - \tilde{v}_{t-1,0} \right\|_2^2 \right] + \frac{2(d_1 + 4)\ell^2\beta^2}{S_{2,x}} \sum_{k=0}^m \mathbb{E}[\|\tilde{u}_{t-1,k}\|_2^2] \\
&\quad + \frac{2(m+1)\mu_1^2(d_1 + 6)^3\ell^2}{S_{2,x}}. \tag{62}
\end{aligned}$$

Similarly, we obtain

$$\begin{aligned}
&\mathbb{E} \left[ \left\| \nabla_y f_{\mu_2}(\tilde{x}_{t-1, \tilde{m}_{t-1}}, \tilde{y}_{t-1, \tilde{m}_{t-1}}) - \tilde{u}_{t-1, \tilde{m}_{t-1}} \right\|_2^2 \right] \\
&\leq \mathbb{E} \left[ \left\| \nabla_y f_{\mu_2}(\tilde{x}_{t-1,0}, \tilde{y}_{t-1,0}) - \tilde{u}_{t-1,0} \right\|_2^2 \right] + \frac{2(d_2 + 4)\ell^2\beta^2}{S_{2,y}} \sum_{k=0}^m \mathbb{E}[\|\tilde{u}_{t-1,k}\|_2^2] \\
&\quad + \frac{2(m+1)\mu_2^2(d_2 + 6)^3\ell^2}{S_{2,y}}. \tag{63}
\end{aligned}$$

Combining eq. (62) and eq. (63) yields

$$\begin{aligned}
\Delta'_t &\leq \tilde{\Delta}'_{t-1,0} + \left( \frac{2(d_1 + 4)\ell^2\beta^2}{S_{2,x}} + \frac{2(d_2 + 4)\ell^2\beta^2}{S_{2,y}} \right) \sum_{k=0}^m \mathbb{E}[\|\tilde{u}_{t-1,k}\|_2^2] \\
&\quad + \frac{2(m+1)\mu_1^2(d_1 + 6)^3\ell^2}{S_{2,x}} + \frac{2(m+1)\mu_2^2(d_2 + 6)^3\ell^2}{S_{2,y}}. \tag{64}
\end{aligned}$$

For  $\tilde{\Delta}'_{t-1,0}$ , we obtain

$$\begin{aligned}
\tilde{\Delta}'_{t-1,0} &= \mathbb{E}[\|\nabla_x f_{\mu_1}(\tilde{x}_{t-1,0}, \tilde{y}_{t-1,0}) - \tilde{v}_{t-1,0}\|_2^2] + \mathbb{E}[\|\nabla_y f_{\mu_2}(\tilde{x}_{t-1,0}, \tilde{y}_{t-1,0}) - \tilde{u}_{t-1,0}\|_2^2] \\
&\stackrel{(i)}{\leq} \mathbb{E}[\|\nabla_x f_{\mu_1}(\tilde{x}_{t-1,-1}, \tilde{y}_{t-1,-1}) - \tilde{v}_{t-1,-1}\|_2^2] + \mathbb{E}[\|\nabla_y f_{\mu_2}(\tilde{x}_{t-1,-1}, \tilde{y}_{t-1,-1}) - \tilde{u}_{t-1,-1}\|_2^2] \\
&\quad + \frac{1}{S_{2,x}} \mathbb{E}[\|G(\tilde{x}_{t,0}, \tilde{y}_{t,0}, \nu_i, \xi_i) - G(\tilde{x}_{t,-1}, \tilde{y}_{t,-1}, \nu_{\mathcal{M}_i}, \xi_i)\|_2^2] \\
&\quad + \frac{1}{S_{2,y}} \mathbb{E}[\|H(\tilde{x}_{t,0}, \tilde{y}_{t,0}, \nu_i, \xi_i) - H(\tilde{x}_{t,-1}, \tilde{y}_{t,-1}, \nu_{\mathcal{M}_i}, \xi_i)\|_2^2]
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(ii)}{\leq} \Delta'_{t-1} + \left( \frac{2(d_1+4)\ell^2\alpha^2}{S_{2,x}} + \frac{2(d_2+4)\ell^2\alpha^2}{S_{2,y}} \right) \mathbb{E}[\|v_{t-1}\|_2^2] \\
&\quad + \frac{2\mu_1^2(d_1+6)^3\ell^2}{S_{2,x}} + \frac{2\mu_2^2(d_2+6)^3\ell^2}{S_{2,y}},
\end{aligned} \tag{65}$$

where (i) follows from Lemma 5 and (ii) follows from Lemma 16. Substituting eq. (65) into eq. (64) yields

$$\begin{aligned}
\Delta'_t &\leq \Delta'_{t-1} + \left( \frac{2(d_1+4)\ell^2\alpha^2}{S_{2,x}} + \frac{2(d_2+4)\ell^2\alpha^2}{S_{2,y}} \right) \mathbb{E}[\|v_{t-1}\|_2^2] \\
&\quad + \left( \frac{2(d_1+4)\ell^2\beta^2}{S_{2,x}} + \frac{2(d_2+4)\ell^2\beta^2}{S_{2,y}} \right) \sum_{k=0}^m \mathbb{E}[\|\tilde{u}_{t-1,k}\|_2^2] \\
&\quad + \frac{2(m+2)\mu_1^2(d_1+6)^3\ell^2}{S_{2,x}} + \frac{2(m+2)\mu_2^2(d_2+6)^3\ell^2}{S_{2,y}} \\
&\stackrel{(i)}{\leq} \Delta'_{t-1} + \left( \frac{2(d_1+4)\ell^2\alpha^2}{S_{2,x}} + \frac{2(d_2+4)\ell^2\alpha^2}{S_{2,y}} \right) \mathbb{E}[\|v_{t-1}\|_2^2] \\
&\quad + \frac{2(m+2)\mu_1^2(d_1+6)^3\ell^2}{S_{2,x}} + \frac{2(m+2)\mu_2^2(d_2+6)^3\ell^2}{S_{2,y}} \\
&\quad + \frac{2\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \left[ \mathbb{E}[\|\tilde{u}_{t,0}\|_2^2] + (m+1) \left( \frac{2\mu_2^2\ell\kappa}{\beta}(d_2+3)^3 + 7\mu_2^2(d_2+6)^3\ell^2 \right) \right].
\end{aligned} \tag{66}$$

where (i) follows from Lemma 20. We next bound the term  $\mathbb{E}[\|\tilde{u}_{t-1,0}\|_2^2]$  as follows:

$$\begin{aligned}
&\mathbb{E}[\|\tilde{u}_{t-1,0}\|_2^2] \\
&= \mathbb{E}[\|\tilde{u}_{t-1,0} - \nabla_y f_{\mu_2}(x_t, y_{t-1}) + \nabla_y f_{\mu_2}(x_t, y_{t-1}) - \nabla_y f_{\mu_2}(x_{t-1}, y_{t-1}) + \nabla_y f_{\mu_2}(x_{t-1}, y_{t-1})\|_2^2] \\
&\leq 3\mathbb{E}[\|\tilde{u}_{t-1,0} - \nabla_y f_{\mu_2}(x_t, y_{t-1})\|_2^2] + 3\mathbb{E}[\|\nabla_y f_{\mu_2}(x_t, y_{t-1}) - \nabla_y f_{\mu_2}(x_{t-1}, y_{t-1})\|_2^2] \\
&\quad + 3\mathbb{E}[\|\nabla_y f_{\mu_2}(x_{t-1}, y_{t-1})\|_2^2] \\
&\stackrel{(i)}{\leq} 3\mathbb{E}[\|\tilde{u}_{t-1,0} - \nabla_y f_{\mu_2}(x_t, y_{t-1})\|_2^2] + 3\ell^2\mathbb{E}[\|x_t - x_{t-1}\|_2^2] + 3\delta'_{t-1} \\
&= 3\mathbb{E}[\|\tilde{u}_{t-1,0} - \nabla_y f_{\mu_2}(\tilde{x}_{t-1,0}, \tilde{y}_{t-1,0})\|_2^2] + 3\alpha^2\ell^2\mathbb{E}[\|v_{t-1}\|_2^2] + 3\delta'_{t-1} \\
&\leq 3\tilde{\Delta}'_{t-1,0} + 3\alpha^2\ell^2\mathbb{E}[\|v_{t-1}\|_2^2] + 3\delta'_{t-1} \\
&\stackrel{(ii)}{\leq} 3\Delta'_{t-1} + 3\delta'_{t-1} + \left[ 3 + \frac{6(d_1+4)}{S_{2,x}} + \frac{6(d_2+4)}{S_{2,y}} \right] \alpha^2\ell^2\mathbb{E}[\|v_{t-1}\|_2^2] + 6\ell^2 \left[ \frac{\mu_1^2(d_1+6)^3}{S_{2,x}} + \frac{\mu_2^2(d_2+6)^3}{S_{2,y}} \right] \\
&\stackrel{(iii)}{\leq} 3\Delta'_{t-1} + 3\delta'_{t-1} + 9\alpha^2\ell^2\mathbb{E}[\|v_{t-1}\|_2^2] + 6\ell^2 \left[ \frac{\mu_1^2(d_1+6)^3}{S_{2,x}} + \frac{\mu_2^2(d_2+6)^3}{S_{2,y}} \right]
\end{aligned} \tag{67}$$

where (i) follows from Lemma 12, and (ii) follows from eq. (65), and (iii) follows from the fact that  $S_{2,x} \geq 2(d_1+4)$  and  $S_{2,y} \geq 2(d_2+4)$ . Substituting eq. (67) into eq. (66) yields

$$\begin{aligned}
\Delta'_t &\leq \left[ 1 + \frac{6\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \right] \Delta'_{t-1} + \frac{6\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \delta'_{t-1} \\
&\quad + 2\ell^2\alpha^2 \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \left( 1 + \frac{9\ell^2\beta^2}{1-b} \right) \mathbb{E}[\|v_{t-1}\|_2^2] \\
&\quad + \frac{2\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \left\{ 6\ell^2 \left[ \frac{\mu_1^2(d_1+6)^3}{S_{2,x}} + \frac{\mu_2^2(d_2+6)^3}{S_{2,y}} \right] + (m+1) \left( \frac{2\mu_2^2\ell\kappa}{\beta}(d_2+3)^3 \right. \right. \\
&\quad \left. \left. + 7\mu_2^2(d_2+6)^3\ell^2 \right) \right\} + \frac{2(m+2)\mu_1^2(d_1+6)^3\ell^2}{S_{2,x}} + \frac{2(m+2)\mu_2^2(d_2+6)^3\ell^2}{S_{2,y}} \\
&\stackrel{(i)}{\leq} \left[ 1 + \frac{6\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \right] \Delta'_{t-1} + \frac{6\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \delta'_{t-1}
\end{aligned}$$

$$+ 2\ell^2\alpha^2 \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \left( 1 + \frac{9\ell^2\beta^2}{1-\beta} \right) \mathbb{E}[\|v_{t-1}\|_2^2] + \pi_\Delta(d_1, d_2, \mu_1, \mu_2). \quad (68)$$

□

**Lemma 24.** Suppose Assumptions 2-3 hold. Let  $S_{2,x} \geq 2d_1 + 8$  and  $S_{2,y} \geq 2d_1 + 8$ . Then, we have

$$\delta'_t \leq \left( \frac{4}{\beta\mu(m+1)} + \frac{3\ell\beta}{2-\ell\beta} \right) \delta'_{t-1} + \frac{2+2\ell\beta}{2-\ell\beta} \Delta'_{t-1} + \left( \frac{4\ell^2\alpha^2}{\beta\mu(m+1)} + 2\ell^2\alpha^2 + \frac{9\ell^3\beta\alpha^2}{2-\ell\beta} \right) \mathbb{E}[\|v_{t-1}\|_2^2] + \pi_\delta(d_1, d_2, \mu_1, \mu_2),$$

where

$$\pi_\delta(d_1, d_2, \mu_1, \mu_2) = \frac{2\ell^2(2+2\ell\beta)}{2-\ell\beta} \left( \frac{\mu_1^2(d_1+6)^3}{S_{2,x}} + \frac{\mu_2^2(d_2+6)^3}{S_{2,y}} \right) + \frac{2}{\beta(m+1)} \left( \frac{\mu_2^2}{4\mu} \ell^2(d_2+3)^3 + \mu_2^2 \ell d_2 \right).$$

Furthermore, if we let  $\beta = \frac{2}{13\ell}$ ,  $m = 104\kappa - 1$ , then we have

$$\pi_\delta(d_1, d_2, \mu_1, \mu_2) = \frac{5}{2} \mu_1^2 \ell^2 (d_1+6)^3 + 3\mu_2^2 \ell^2 (d_2+6)^3 + \frac{1}{8} \mu_2^2 \mu \ell d_2.$$

*Proof.* Using the result in Lemma 19, and recalling in Algorithm 4 that  $\nabla g_{t,\mu_2}(\tilde{y}_t, \tilde{m}_t) = \nabla_y f(x_t, y_t)$  and  $\nabla g_{t,\mu_2}(\tilde{y}_t, 0) = \nabla_y f_{\mu_2}(x_{t+1}, y_t)$ , we have

$$\begin{aligned} \delta'_{t+1} &\leq \frac{2}{\beta\mu(m+1)} \mathbb{E}[\|\nabla_y f_{\mu_2}(x_{t+1}, y_t)\|_2^2] + \mathbb{E}[\|\nabla g_{t,\mu_2}(\tilde{y}_t, 0) - \tilde{u}_{t,0}\|_2^2] + \frac{\ell\beta}{2-\ell\beta} \mathbb{E}[\|\tilde{u}_{t,0}\|_2^2] \\ &\quad + \frac{2}{\beta(m+1)} \left( \frac{\mu_2^2}{4\mu} \ell^2(d_2+3)^3 + \mu_2^2 \ell d_2 \right) \\ &\leq \frac{2}{\beta\mu(m+1)} \mathbb{E}[\|\nabla_y f_{\mu_2}(x_{t+1}, y_t)\|_2^2] + \tilde{\Delta}'_{t,0} + \frac{\ell\beta}{2-\ell\beta} \mathbb{E}[\|\tilde{u}_{t,0}\|_2^2] \\ &\quad + \frac{2}{\beta(m+1)} \left( \frac{\mu_2^2}{4\mu} \ell^2(d_2+3)^3 + \mu_2^2 \ell d_2 \right) \\ &\leq \frac{4}{\beta\mu(m+1)} \mathbb{E}[\|\nabla_y f_{\mu_2}(x_{t+1}, y_t) - \nabla_y f_{\mu_2}(x_t, y_t)\|_2^2] + \frac{4}{\beta\mu(m+1)} \mathbb{E}[\|\nabla_y f_{\mu_2}(x_t, y_t)\|_2^2] \\ &\quad + \tilde{\Delta}'_{t,0} + \frac{\ell\beta}{2-\ell\beta} \mathbb{E}[\|\tilde{u}_{t,0}\|_2^2] + \frac{2}{\beta(m+1)} \left( \frac{\mu_2^2}{4\mu} \ell^2(d_2+3)^3 + \mu_2^2 \ell d_2 \right) \\ &\leq \frac{4\ell^2\alpha^2}{\beta\mu(m+1)} \mathbb{E}[\|v_t\|_2^2] + \frac{4}{\beta\mu(m+1)} \delta'_t + \tilde{\Delta}'_{t,0} + \frac{\ell\beta}{2-\ell\beta} \mathbb{E}[\|\tilde{u}_{t,0}\|_2^2] \\ &\quad + \frac{2}{\beta(m+1)} \left( \frac{\mu_2^2}{4\mu} \ell^2(d_2+3)^3 + \mu_2^2 \ell d_2 \right) \\ &\stackrel{(i)}{\leq} \frac{4\ell^2\alpha^2}{\beta\mu(m+1)} \mathbb{E}[\|v_t\|_2^2] + \frac{4}{\beta\mu(m+1)} \delta'_t \\ &\quad + \Delta'_t + 2\ell^2\alpha^2 \mathbb{E}[\|v_t\|_2^2] + 2\ell^2 \left( \frac{\mu_1^2(d_1+6)^3}{S_{2,x}} + \frac{\mu_2^2(d_2+6)^3}{S_{2,y}} \right) \\ &\quad + \frac{\ell\beta}{2-\ell\beta} \left[ 3\Delta'_t + 3\delta'_t + 9\ell^2\alpha^2 \mathbb{E}[\|v_t\|_2^2] + 6\ell^2 \left( \frac{\mu_1^2(d_1+6)^3}{S_{2,x}} + \frac{\mu_2^2(d_2+6)^3}{S_{2,y}} \right) \right] \\ &\quad + \frac{2}{\beta(m+1)} \left( \frac{\mu_2^2}{4\mu} \ell^2(d_2+3)^3 + \mu_2^2 \ell d_2 \right) \\ &= \left( \frac{4}{\beta\mu(m+1)} + \frac{3\ell\beta}{2-\ell\beta} \right) \delta'_t + \frac{2+2\ell\beta}{2-\ell\beta} \Delta'_t + \left( \frac{4\ell^2\alpha^2}{\beta\mu(m+1)} + 2\ell^2\alpha^2 + \frac{9\ell^3\beta\alpha^2}{2-\ell\beta} \right) \mathbb{E}[\|v_t\|_2^2] \\ &\quad + \frac{2\ell^2(2+2\ell\beta)}{2-\ell\beta} \left( \frac{\mu_1^2(d_1+6)^3}{S_{2,x}} + \frac{\mu_2^2(d_2+6)^3}{S_{2,y}} \right) + \frac{2}{\beta(m+1)} \left( \frac{\mu_2^2}{4\mu} \ell^2(d_2+3)^3 + \mu_2^2 \ell d_2 \right), \\ &\leq \left( \frac{4}{\beta\mu(m+1)} + \frac{3\ell\beta}{2-\ell\beta} \right) \delta'_t + \frac{2+2\ell\beta}{2-\ell\beta} \Delta'_t + \left( \frac{4\ell^2\alpha^2}{\beta\mu(m+1)} + 2\ell^2\alpha^2 + \frac{9\ell^3\beta\alpha^2}{2-\ell\beta} \right) \mathbb{E}[\|v_t\|_2^2] \end{aligned}$$

$$+ \pi_\delta(d_1, d_2, \mu_1, \mu_2), \quad (69)$$

where (i) follows from eq. (65) and eq. (67), and from the fact that  $S_{2,x} \geq 2d_1 + 8$  and  $S_{2,y} \geq 2d_2 + 8$ . The proof is complete by shifting the index in eq. (69) from  $t$  to  $t - 1$ .  $\square$

We restate Theorem 2 as follows to include the specifics of the parameters.

**Theorem 4** (Restate of Theorem 2 with parameter specifics). *Let Assumptions 1, 2, 4, and 3 hold and apply ZO-SREDA-Boost in Algorithm 4 to solve the problem in eq. (1) with the following parameters:*

$$\begin{aligned} \zeta &= \frac{1}{\kappa}, \quad \alpha = \frac{1}{24(\kappa + 1)\ell}, \quad \beta = \frac{2}{13\ell}, \quad q = \frac{2800\kappa}{13\epsilon(\kappa + 1)}, \\ m &= 104\kappa - 1, \quad S_{2,x} = \frac{5600(d_1 + 4)\kappa}{\epsilon}, \quad S_{2,y} = \frac{5600(d_2 + 4)\kappa}{\epsilon}, \\ S_1 &= \frac{40320\sigma^2\kappa^2}{\epsilon^2}, \quad T = \max\{1728(\kappa + 1)\ell \frac{\Phi(x_0) - \Phi^*}{\epsilon^2}, \frac{810\kappa}{\epsilon^2}\}, \\ \delta &= \frac{\epsilon}{71\kappa\ell\sqrt{d_1 + d_2}}, \quad \mu_1 = \frac{\epsilon}{71\kappa^{2.5}\ell(d_1 + 6)^{1.5}}, \quad \mu_2 = \frac{\epsilon}{71\kappa^{2.5}\ell(d_2 + 6)^{1.5}}. \end{aligned}$$

Algorithm 4 outputs  $\hat{x}$  such that

$$\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2] \leq \epsilon$$

with at most  $\mathcal{O}((d_1 + d_2)\kappa^3\epsilon^{-3})$  function queries.

*Proof.* Recalling from eq. (18), we have

$$\begin{aligned} \mathbb{E}[\Phi(x_{t+1})] &\leq \mathbb{E}[\Phi(x_t)] + \alpha\kappa^2\delta_t + \alpha\Delta_t - \left(\frac{\alpha}{2} - \frac{L\alpha^2}{2}\right) \mathbb{E}[\|v_t\|_2^2] \\ &\stackrel{(i)}{\leq} \mathbb{E}[\Phi(x_t)] + 2\alpha\kappa^2\delta'_t + 2\alpha\Delta'_t - \left(\frac{\alpha}{2} - \frac{L\alpha^2}{2}\right) \mathbb{E}[\|v_t\|_2^2] \\ &\quad + \frac{\mu_2\alpha(\kappa^2 + 1)}{2}\ell^2(d_2 + 3)^3 + \frac{\mu_1\alpha}{2}\ell^2(d_1 + 3)^3, \end{aligned} \quad (70)$$

where (i) follows from Lemma 22. Rearranging eq. (70) and taking the summation over  $t = \{0, 1, \dots, T - 1\}$  yield

$$\begin{aligned} \left(\frac{\alpha}{2} - \frac{L\alpha^2}{2}\right) \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] &\leq \Phi(x_0) - \mathbb{E}[\Phi(x_T)] + 2\alpha\kappa^2 \sum_{t=0}^{T-1} \delta'_t + 2\alpha \sum_{t=0}^{T-1} \Delta'_t \\ &\quad + \alpha T\pi(d_1, d_2, \mu_1, \mu_2). \end{aligned} \quad (71)$$

Note that in eq. (71) we define

$$\pi(d_1, d_2, \mu_1, \mu_2) = \frac{\mu_2^2(\kappa^2 + 1)}{2}\ell^2(d_2 + 3)^3 + \frac{\mu_1^2}{2}\ell^2(d_1 + 3)^3. \quad (72)$$

Then we proceed to prove Theorem 2/Theorem 4 in the following five steps.

**Step 1.** We establish the induction relationships for the tracking error and gradient estimation error with respect to the Gaussian smoothed function upon one outer-loop update for SREDA-Boost. Namely, we develop the relationship between  $\delta'_t$  and  $\delta'_{t-1}$  as well as that between  $\Delta'_t$  and  $\Delta'_{t-1}$ , which are captured in Lemma 23 and Lemma 24.

**Step 2.** Based on Step 1, we provide the bounds on the inter-related accumulative errors  $\sum_{t=0}^{T-1} \Delta'_t$  and  $\sum_{t=0}^{T-1} \delta'_t$  over the entire execution of the algorithm.

We first consider  $\sum_{t=0}^{T-1} \Delta'_t$ , for any  $(n_T - 1)q \leq t' < T - 1$ . Applying the inequality in Lemma 23 recursively, we obtain the following bound

$$\Delta'_t \leq \left[1 + \frac{6\ell^2\beta^2}{1-b} \left(\frac{d_1 + 4}{S_{2,x}} + \frac{d_2 + 4}{S_{2,y}}\right)\right] \Delta'_{t-1} + \frac{6\ell^2\beta^2}{1-b} \left(\frac{d_1 + 4}{S_{2,x}} + \frac{d_2 + 4}{S_{2,y}}\right) \delta'_{t-1}$$

$$\begin{aligned}
& + 2\ell^2\alpha^2 \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \left( 1 + \frac{9\ell^2\beta^2}{1-b} \right) \mathbb{E}[\|v_{T-2}\|_2^2] + \pi_\Delta(d_1, d_2, \mu_1, \mu_2, S_2) \\
& \leq \left[ 1 + \frac{6\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \right]^{t-t'} \Delta'_{t'} \\
& \quad + \frac{6\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \sum_{p=t'}^{t-1} \left[ 1 + \frac{6\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \right]^{p-t'} \delta'_p \\
& \quad + 2\ell^2\alpha^2 \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \left( 1 + \frac{9\ell^2\beta^2}{1-b} \right) \sum_{p=t'}^{t-1} \left[ 1 + \frac{6\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \right]^{p-t'} \mathbb{E}[\|v_t\|_2^2] \\
& \quad + \pi_\Delta(d_1, d_2, \mu_1, \mu_2, S_2) \sum_{p=t'}^{t-1} \left[ 1 + \frac{6\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \right]^{p-t'} \\
& \stackrel{(i)}{\leq} 2\Delta'_{t'} + \frac{6\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \sum_{p=t'}^{t-1} \delta'_p \\
& \quad + 2\ell^2\alpha^2 \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \left( 1 + \frac{9\ell^2\beta^2}{1-b} \right) \sum_{p=t'}^{t-1} \mathbb{E}[\|v_t\|_2^2] \\
& \quad + 2\pi_\Delta(d_1, d_2, \mu_1, \mu_2, S_2), \tag{73}
\end{aligned}$$

where (i) follows from the fact that

$$\begin{aligned}
& \left[ 1 + \frac{6\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \right]^{p-t'} \leq \left[ 1 + \frac{6\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \right]^q \\
& \stackrel{(ii)}{\leq} 1 + \frac{\frac{6q\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right)}{1 - \frac{6(q-1)\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right)} \stackrel{(iii)}{\leq} 2,
\end{aligned}$$

where (ii) follows from the Bernoulli's inequality Li & Yeh (2013) (eq. (21)), and (iii) follows from the fact that  $q = (1-b) \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right)^{-1}$ ,  $\beta = \frac{2}{13\ell}$ ,  $\left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) < 1$ , and  $b = 1 - \frac{\beta\mu\ell}{2(\mu+\ell)}$ , which further implies that

$$\frac{\frac{6q\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right)}{1 - \frac{6(q-1)\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right)} \leq \frac{\frac{6q\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right)}{1 - \frac{6q\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right)} = \frac{6\ell^2\beta^2}{1 - 6\ell^2\beta^2} < 1.$$

Letting  $t' = (n_T - 1)q$  and taking summation of eq. (73) over  $t = \{(n_T - 1)q, \dots, T - 1\}$  yield

$$\begin{aligned}
& \sum_{t=(n_T-1)q}^{T-1} \Delta'_t \leq 2(T - (n_T - 1)q) \Delta'_{(n_T-1)q} + \frac{6\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \sum_{t=(n_T-1)q}^{T-1} \sum_{p=(n_T-1)q}^{t-1} \delta'_p \\
& \quad + 2\ell^2\alpha^2 \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \left( 1 + \frac{9\ell^2\beta^2}{1-b} \right) \sum_{t=(n_T-1)q}^{T-1} \sum_{p=(n_T-1)q}^{t-1} \mathbb{E}[\|v_p\|_2^2] \\
& \quad + 2(T - (n_T - 1)q) \pi_\Delta(d_1, d_2, \mu_1, \mu_2, S_2) \\
& \stackrel{(i)}{\leq} 2(T - (n_T - 1)q) \epsilon(S_1, \delta) + \frac{6q\ell^2\beta^2}{1-b} \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \sum_{t=(n_T-1)q}^{T-2} \delta'_t \\
& \quad + 2q\ell^2\alpha^2 \left( \frac{d_1+4}{S_{2,x}} + \frac{d_2+4}{S_{2,y}} \right) \left( 1 + \frac{9\ell^2\beta^2}{1-b} \right) \sum_{t=(n_T-1)q}^{T-2} \mathbb{E}[\|v_t\|_2^2] \\
& \quad + 2(T - (n_T - 1)q) \pi_\Delta(d_1, d_2, \mu_1, \mu_2)
\end{aligned}$$



$$\begin{aligned}
&= 2(T - (n_T - 1)q)\epsilon(S_1, \delta) + 6\ell^2\beta^2 \sum_{t=(n_T-1)q}^{T-2} \delta'_t \\
&\quad + 2\ell^2\alpha^2(1-b) \left(1 + \frac{9\ell^2\beta^2}{1-b}\right) \sum_{t=(n_T-1)q}^{T-2} \mathbb{E}[\|v_t\|_2^2] \\
&\quad + 2(T - (n_T - 1)q)\pi_\Delta(d_1, d_2, \mu_1, \mu_2) \\
&\leq 2(T - (n_T - 1)q)\epsilon(S_1, \delta) + 6\ell^2\beta^2 \sum_{t=(n_T-1)q}^{T-2} \delta'_t + 2\ell^2\alpha^2(1+9\ell^2\beta^2) \sum_{t=(n_T-1)q}^{T-2} \mathbb{E}[\|v_t\|_2^2] \\
&\quad + 2(T - (n_T - 1)q)\pi_\Delta(d_1, d_2, \mu_1, \mu_2) \\
&\stackrel{(ii)}{\leq} 2(T - (n_T - 1)q)\epsilon(S_1, \delta) + \frac{1}{7} \sum_{t=(n_T-1)q}^{T-2} \delta'_t + 3\ell^2\alpha^2 \sum_{t=(n_T-1)q}^{T-2} \mathbb{E}[\|v_t\|_2^2] \\
&\quad + 2(T - (n_T - 1)q)\pi_\Delta(d_1, d_2, \mu_1, \mu_2), \tag{74}
\end{aligned}$$

where (i) follows from the fact that  $\Delta'_{(n_T-n)q} \leq \epsilon(S_1, \delta)$  for all  $n \leq n_T$  (following from Lemma 4) and the definition of  $\epsilon(S_1, \delta)$  in Lemma 15,

$$\sum_{t=(n_T-1)q}^{T-1} \sum_{p=(n_T-1)q}^{t-1} \delta'_p \leq q \sum_{t=(n_T-1)q}^{T-2} \delta'_t,$$

and

$$\sum_{t=(n_T-1)q}^{T-1} \sum_{p=(n_T-1)q}^{t-1} \mathbb{E}[\|v_t\|_2^2] \leq q \sum_{t=(n_T-1)q}^{T-2} \mathbb{E}[\|v_t\|_2^2],$$

and in (ii) we use the fact that  $\beta = \frac{2}{13\ell}$ . Applying steps similar to those in eq. (22) for iterations over  $t = \{(n_T - n_t)q, \dots, (n_T - n_t + 1)q - 1\}$  yields

$$\begin{aligned}
\sum_{t=(n_T-n_t)q}^{(n_T-n_t+1)q-1} \Delta'_t &\leq 2q\epsilon(S_1, \delta) + \frac{1}{7} \sum_{t=(n_T-n_t)q}^{(n_T-n_t+1)q-1} \delta'_t + 3\ell^2\alpha^2 \sum_{t=(n_T-n_t)q}^{(n_T-n_t+1)q-1} \mathbb{E}[\|v_t\|_2^2] \\
&\quad + 2q\pi_\Delta(d_1, d_2, \mu_1, \mu_2). \tag{75}
\end{aligned}$$

Taking summation of eq. (75) over  $n = \{2, \dots, n_T\}$  and combining with eq. (74) yield

$$\sum_{t=0}^{T-1} \Delta'_t \leq 2T\epsilon(S_1, \delta) + \frac{1}{7} \sum_{t=0}^{T-1} \delta'_t + 3\ell^2\alpha^2 \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] + 2T\pi_\Delta(d_1, d_2, \mu_1, \mu_2). \tag{76}$$

Then we consider the upper bound on  $\sum_{t=0}^{T-1} \delta'_t$ . Since  $m = \frac{16}{\mu\beta} - 1$  and  $\beta = \frac{2}{13\ell}$ , Lemma 24 implies

$$\delta'_t \leq \frac{1}{2}\delta'_{t-1} + \frac{5}{4}\Delta'_{t-1} + 3\ell^2\alpha^2\mathbb{E}[\|v_{t-1}\|_2^2] + \pi_\delta(d_1, d_2, \mu_1, \mu_2). \tag{77}$$

Applying eq. (77) recursively from  $t$  to 0 yields

$$\delta'_t \leq \frac{1}{2^t}\delta'_0 + \frac{5}{4} \sum_{p=0}^{t-1} \frac{1}{2^p} \Delta'_p + 3\ell^2\alpha^2 \sum_{p=0}^{t-1} \frac{1}{2^p} \mathbb{E}[\|v_p\|_2^2] + \pi_\delta(d_1, d_2, \mu_1, \mu_2) \sum_{p=0}^{t-1} \frac{1}{2^p}. \tag{78}$$

Taking the summation of eq. (78) over  $t = \{0, 1, \dots, T-1\}$  yields

$$\begin{aligned}
\sum_{t=0}^{T-1} \delta'_t &\leq \delta'_0 \sum_{t=0}^{T-1} \frac{1}{2^t} + \frac{5}{4} \sum_{t=0}^{T-1} \sum_{p=0}^{t-1} \frac{1}{2^p} \Delta'_p + 3\ell^2\alpha^2 \sum_{t=0}^{T-1} \sum_{p=0}^{t-1} \frac{1}{2^p} \mathbb{E}[\|v_p\|_2^2] \\
&\quad + \pi_\delta(d_1, d_2, \mu_1, \mu_2) \sum_{t=0}^{T-1} \sum_{p=0}^{t-1} \frac{1}{2^p}
\end{aligned}$$

$$\leq 2\delta'_0 + \frac{5}{2} \sum_{t=0}^{T-2} \Delta'_t + 6\ell^2 \alpha^2 \sum_{t=0}^{T-2} \mathbb{E}[\|v_t\|_2^2] + 2T\pi_\delta(d_1, d_2, \mu_1, \mu_2). \quad (79)$$

**Step 3.** We decouple the bounds on  $\sum_{t=0}^{T-1} \Delta'_t$  and  $\sum_{t=0}^{T-1} \delta'_t$  in Step 2 from each other, and establish their separate relationships with the accumulative gradient estimators  $\sum_{i=0}^{T-1} \mathbb{E}[\|v_t\|_2^2]$ .

Substituting eq. (79) into eq. (76) yields

$$\begin{aligned} \sum_{t=0}^{T-1} \Delta'_t &\leq 2T\epsilon(S_1, \delta) + \frac{2}{7}\delta'_0 + 4\alpha^2\ell^2 \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] + \frac{5}{14} \sum_{t=0}^{T-2} \Delta'_t \\ &\quad + 2T\pi_\Delta(d_1, d_2, \mu_1, \mu_2) + \frac{2}{7}T\pi_\delta(d_1, d_2, \mu_1, \mu_2), \end{aligned}$$

which implies

$$\begin{aligned} \sum_{t=0}^{T-1} \Delta'_t &\leq 4T\epsilon(S_1, \delta) + \frac{1}{2}\delta'_0 + 7\alpha^2\ell^2 \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \\ &\quad + \frac{1}{2}T\pi_\Delta(d_1, d_2, \mu_1, \mu_2) + 4T\pi_\delta(d_1, d_2, \mu_1, \mu_2). \end{aligned} \quad (80)$$

Substituting eq. (80) into eq. (79) yields

$$\begin{aligned} \sum_{t=0}^{T-1} \delta'_t &\leq 10T\epsilon(S_1, \delta) + 4\delta'_0 + 24\alpha^2\ell^2 \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \\ &\quad + 10T\pi_\Delta(d_1, d_2, \mu_1, \mu_2) + 4T\pi_\delta(d_1, d_2, \mu_1, \mu_2). \end{aligned} \quad (81)$$

**Step 4.** We bound  $\sum_{i=0}^{T-1} \mathbb{E}[\|v_t\|_2^2]$ , and further cancel out the impact of  $\sum_{t=0}^{T-1} \Delta'_t$  and  $\sum_{t=0}^{T-1} \delta'_t$  by exploiting Step 3.

Substituting eq. (80) and eq. (81) into eq. (71) yields

$$\begin{aligned} &\left(\frac{\alpha}{2} - \frac{L\alpha^2}{2}\right) \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \\ &\leq \Phi(x_0) - \mathbb{E}[\Phi(x_T)] + (20\kappa^2 + 8)\alpha T\epsilon(S_1, \delta) + (8\kappa^2 + 1)\alpha\delta'_0 + (48\kappa^2 + 14)\alpha^3\ell^2 \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \\ &\quad + (20\kappa^2 + 1)\alpha T\pi_\Delta(d_1, d_2, \mu_1, \mu_2) + (8\kappa^2 + 8)\alpha T\pi_\delta(d_1, d_2, \mu_1, \mu_2) + \alpha T\pi(d_1, d_2, \mu_1, \mu_2) \\ &\stackrel{(i)}{\leq} \Phi(x_0) - \mathbb{E}[\Phi(x_T)] + 28\kappa^2\alpha T\epsilon(S_1, \delta) + 9\kappa^2\alpha\delta'_0 + 62\alpha^3L^2 \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \\ &\quad + 21\kappa^2\alpha T\pi_\Delta(d_1, d_2, \mu_1, \mu_2) + 16\kappa^2\alpha T\pi_\delta(d_1, d_2, \mu_1, \mu_2) + \alpha T\pi(d_1, d_2, \mu_1, \mu_2), \end{aligned} \quad (82)$$

where (i) follows from the fact that  $L = (1 + \kappa)\ell$  and  $\kappa > 1$ . Rearranging eq. (82), we have

$$\begin{aligned} &\left(\frac{\alpha}{2} - \frac{L\alpha^2}{2} - 62L^2\alpha^3\right) \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \\ &\leq \Phi(x_0) - \mathbb{E}[\Phi(x_T)] + 28\kappa^2\alpha T\epsilon(S_1, \delta) + 9\kappa^2\alpha\delta'_0 \\ &\quad + 21\kappa^2\alpha T\pi_\Delta(d_1, d_2, \mu_1, \mu_2) + 16\kappa^2\alpha T\pi_\delta(d_1, d_2, \mu_1, \mu_2) + \alpha T\pi(d_1, d_2, \mu_1, \mu_2). \end{aligned} \quad (83)$$

Since  $\alpha = \frac{1}{24L}$ , we obtain

$$\frac{\alpha}{2} - \frac{L\alpha^2}{2} - 62L^2\alpha^3 = \frac{214}{13824L} \geq \frac{1}{72L}. \quad (84)$$

Substituting eq. (84) into eq. (83) and applying Assumption 1 yield

$$\sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \leq 72L(\Phi(x_0) - \Phi^*) + 84\kappa^2T\epsilon(S_1, \delta) + 27\kappa^2\delta'_0$$

$$\begin{aligned}
& + 63\kappa^2 T \pi_{\Delta}(d_1, d_2, \mu_1, \mu_2) + 48\kappa^2 T \pi_{\delta}(d_1, d_2, \mu_1, \mu_2) \\
& + 3T\pi(d_1, d_2, \mu_1, \mu_2).
\end{aligned} \tag{85}$$

**Step 5.** We establish the convergence bound on  $\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2]$  based on the bounds on its estimators  $\sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2]$  and the two error bounds  $\sum_{t=0}^{T-1} \Delta'_t$ , and  $\sum_{t=0}^{T-1} \delta'_t$ .

Recall eq. (34) we have

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\Phi(x_t)\|_2^2] & \leq 3\kappa^2 \sum_{t=0}^{T-1} \delta_t + 3 \sum_{t=0}^{T-1} \Delta_t + 3 \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \\
& \stackrel{(i)}{\leq} 6\kappa^2 \sum_{t=0}^{T-1} \delta'_t + 6 \sum_{t=0}^{T-1} \Delta'_t + 3 \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] + 3T\pi(d_1, d_2, \mu_1, \mu_2)
\end{aligned} \tag{86}$$

where (i) follows from Lemma 22. Substituting eq. (80), eq. (81) and eq. (85) into eq. (86) yields

$$\begin{aligned}
& \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\Phi(x_t)\|_2^2] \\
& \leq (60\kappa^2 + 24)T\epsilon(S_1, \delta) + (24\kappa^2 + 3)\delta'_0 + (60\kappa^2 + 3)T\pi_{\Delta}(d_1, d_2, \mu_1, \mu_2) \\
& \quad + (24\kappa^2 + 24)T\pi_{\delta}(d_1, d_2, \mu_1, \mu_2) + (144\kappa^2\alpha^2\ell^2 + 42\alpha^2\ell^2 + 3) \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \\
& \quad + 3T\pi(d_1, d_2, \mu_1, \mu_2) \\
& \stackrel{(i)}{\leq} 84\kappa^2 T\epsilon(S_1, \delta) + 27\kappa^2 \delta'_0 + 63\kappa^2 T\pi_{\Delta}(d_1, d_2, \mu_1, \mu_2) + 48\kappa^2 T\pi_{\delta}(d_1, d_2, \mu_1, \mu_2) \\
& \quad + 4 \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] + 3T\pi(d_1, d_2, \mu_1, \mu_2) \\
& \stackrel{(ii)}{\leq} 288L(\Phi(x_0) - \Phi^*) + 420\kappa^2 T\epsilon(S_1, \delta) + 135\kappa^2 \delta'_0 + 315\kappa^2 T\pi_{\Delta}(d_1, d_2, \mu_1, \mu_2) \\
& \quad + 240\kappa^2 T\pi_{\delta}(d_1, d_2, \mu_1, \mu_2) + 15T\pi(d_1, d_2, \mu_1, \mu_2).
\end{aligned} \tag{87}$$

where (i) follows from the fact that  $\kappa > 1$ ,  $L = (\kappa + 1)\ell$  and  $\alpha = \frac{1}{24L}$ , and (ii) follows from eq. (85). Recall  $L = (1 + \kappa)\ell$ . Then, eq. (87) implies that

$$\begin{aligned}
& \mathbb{E}[\|\nabla\Phi(\hat{x})\|_2^2] \\
& \leq 288(\kappa + 1)\ell \frac{\Phi(x_0) - \Phi^*}{T} + 420\kappa^2 \epsilon(S_1, \delta) + \frac{135\kappa^2 \delta'_0}{T} \\
& \quad + 315\kappa^2 \pi_{\Delta}(d_1, d_2, \mu_1, \mu_2) + 240\kappa^2 \pi_{\delta}(d_1, d_2, \mu_1, \mu_2) + 15\pi(d_1, d_2, \mu_1, \mu_2).
\end{aligned} \tag{88}$$

Recalling Lemma 15, we have

$$\epsilon(S_1, \delta) \leq \frac{(d_1 + d_2)\ell^2\delta^2}{2} + \frac{4\sigma^2}{S_1} + \frac{\mu_1^2}{2}\ell^2(d_1 + 3)^3 + \frac{\mu_2^2}{2}\ell^2(d_2 + 3)^3.$$

If we let  $\delta'_0 \leq \frac{1}{\kappa}$ ,  $T = \max\{1728(\kappa + 1)\ell \frac{\Phi(x_0) - \Phi^*}{\epsilon^2}, \frac{810\kappa}{\epsilon^2}\}$ ,  $S_1 = \frac{40320\sigma^2\kappa^2}{\epsilon^2}$ , and further let  $\delta = \frac{\epsilon}{71\kappa\ell\sqrt{d_1 + d_2}}$ ,  $\mu_1 = \frac{\epsilon}{71\kappa^{2.5}\ell(d_1 + 6)^{1.5}}$  and  $\mu_2 = \frac{\epsilon}{71\kappa^{2.5}\ell(d_2 + 6)^{1.5}}$ , according to the definition of  $\epsilon(S_1, \delta)$  (Lemma 15),  $\pi_{\Delta}(d_1, d_2, \mu_1, \mu_2)$  (Lemma 23),  $\pi_{\delta}(d_1, d_2, \mu_1, \mu_2)$  (Lemma 24) and  $\pi(d_1, d_2, \mu_1, \mu_2)$  (eq. (72)), then we have  $420\kappa^2 \epsilon(S_1, \delta) \leq \frac{\epsilon^2}{6}$ , and

$$315\kappa^2 \pi_{\Delta}(d_1, d_2, \mu_1, \mu_2) + 240\kappa^2 \pi_{\delta}(d_1, d_2, \mu_1, \mu_2) + 15\pi(d_1, d_2, \mu_1, \mu_2) \leq \frac{\epsilon^2}{2},$$

which implies

$$\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2] \leq \sqrt{\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2^2]} \leq \epsilon.$$

We also let  $S_{2,x} = \frac{5600(d_1+4)\kappa}{\epsilon}$ ,  $S_{2,y} = \frac{5600(d_2+4)\kappa}{\epsilon}$  and  $q = \frac{2800\kappa}{13\epsilon(\kappa+1)}$ . Then, the total sample complexity is given by

$$\begin{aligned} & T \cdot (S_{2,x} + S_{2,y}) \cdot m + \left\lceil \frac{T}{q} \right\rceil \cdot S_1 \cdot (d_1 + d_2) + T_0 \\ & \leq \Theta \left( \frac{\kappa}{\epsilon^2} \cdot \frac{(d_1 + d_2)\kappa}{\epsilon} \cdot \kappa \right) + \Theta \left( \frac{\kappa}{\epsilon} \cdot \frac{\kappa^2}{\epsilon^2} \cdot (d_1 + d_2) \right) + \Theta(d_2 \kappa \log(\kappa)) \\ & = \mathcal{O} \left( \frac{(d_1 + d_2)\kappa^3}{\epsilon^3} \right), \end{aligned}$$

which completes the proof.  $\square$

## C.6 PROOF OF COROLLARY 2

In the finite-sum case, recall that

$$f(x, y) \triangleq \frac{1}{n} \sum_{i=1}^n F(x, y; \xi_i).$$

Here we modify Algorithm 5 by replacing the mini-batch update used in line 6 of Algorithm 4 with the following update using all samples:

$$\begin{aligned} v_t &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{d_1} \frac{F(x_t + \delta e_j, y_t, \xi_i) - F(x_t - \delta e_j, y_t, \xi_i)}{2\delta} e_j, \\ u_t &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{d_2} \frac{F(x_t, y_t + \delta e_j, \xi_i) - F(x_t, y_t - \delta e_j, \xi_i)}{2\delta} e_j, \end{aligned}$$

where  $e_j$  denotes the  $j$ -th canonical unit basis vector. In this case, if  $\text{mod}(k, q) = 0$ , then we have

$$\epsilon(S_1, \delta) \leq \frac{(d_1 + d_2)\ell^2\delta^2}{2} + \frac{\mu_1^2}{2}\ell^2(d_1 + 3)^3 + \frac{\mu_2^2}{2}\ell^2(d_2 + 3)^3. \quad (89)$$

**Case 1:**  $n \geq \kappa^2$

Substituting eq. (89) into eq. (88), it can be checked easily that under the same parameter settings for  $\delta'_0$ ,  $T$ ,  $\delta$ ,  $\mu_1$  and  $\mu_2$  in Theorem 2, we have

$$\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2] \leq \sqrt{\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2^2]} \leq \epsilon.$$

Then, let  $S_{2,x} = 5600(d_1 + 4)\kappa\sqrt{n}$ ,  $S_{2,y} = 5600(d_2 + 4)\kappa\sqrt{n}$  and  $q = \frac{2800\kappa\sqrt{n}}{13(\kappa+1)}$ . Recalling the sample complexity result of ZO-iSARSH in the finite-sum case in Appendix C.4, we have  $T_0 = \mathcal{O}(d_2(\kappa + n) \log(\kappa))$ . The total sample complexity is given by

$$\begin{aligned} & T \cdot (S_{2,x} + S_{2,y}) \cdot m + \left\lceil \frac{T}{q} \right\rceil \cdot S_1 \cdot (d_1 + d_2) + T_0 \\ & \leq \Theta \left( \frac{\kappa}{\epsilon^2} \cdot (d_1 + d_2)\sqrt{n} \cdot \kappa \right) + \Theta \left( \left\lceil \frac{\kappa^2}{\sqrt{n}\epsilon^2} \right\rceil \cdot n \cdot (d_1 + d_2) \right) + \Theta(d_2(\kappa + n)\kappa \log(\kappa)) \\ & = \mathcal{O}((d_1 + d_2)(\sqrt{n}\kappa^2\epsilon^{-2} + n)) + \mathcal{O}(d_2(\kappa^2 + \kappa n) \log(\kappa)). \end{aligned}$$

**Case 2:**  $n \leq \kappa^2$

In this case, we let  $S_{2,x} = 56(d_1 + 4) + 420$ ,  $S_{2,y} = 56(d_2 + 4) + 420$  and  $q = 1$ . Then we have

$$\Delta'_t \leq \epsilon_\Delta = \frac{(d_1 + d_2)\ell^2\delta^2}{2} + \frac{\mu_1^2}{2}\ell^2(d_1 + 3)^3 + \frac{\mu_2^2}{2}\ell^2(d_2 + 3)^3, \quad \text{for all } 0 \leq t \leq T-1. \quad (90)$$

Given the value of  $S_{2,x}$  and  $S_{2,y}$ , it can be checked that the proofs of Lemma 20 and Lemma 24 still hold. Following from the steps similar to those from eq. (71) to eq. (79), we obtain

$$\sum_{t=0}^{T-1} \delta'_t \leq 2\delta'_0 + \frac{5}{2}T\epsilon_\Delta + 6\ell^2\alpha^2 \sum_{t=0}^{T-2} \mathbb{E}[\|v_t\|_2^2] + 2T\pi_\delta(d_1, d_2, \mu_1, \mu_2). \quad (91)$$

Substituting eq. (90) and eq. (91) into eq. (71) yields

$$\begin{aligned} & \left( \frac{\alpha}{2} - \frac{L\alpha^2}{2} \right) \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \\ & \stackrel{(i)}{\leq} \Phi(x_0) - \mathbb{E}[\Phi(x_T)] + 4\alpha\kappa^2\delta'_0 + 7\alpha\kappa^2T\epsilon_\Delta + 12L^2\alpha^3 \sum_{t=0}^{T-2} \mathbb{E}[\|v_t\|_2^2] \\ & \quad + 4\alpha\kappa^2T\pi_\delta(d_1, d_2, \mu_1, \mu_2) + \alpha T\pi(d_1, d_2, \mu_1, \mu_2), \end{aligned} \quad (92)$$

where in (i) we use  $L = (1 + \kappa)\ell$ . Rearranging eq. (92) yields

$$\begin{aligned} & \left( \frac{\alpha}{2} - \frac{L\alpha^2}{2} - 12L^2\alpha^3 \right) \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] \\ & \leq \Phi(x_0) - \mathbb{E}[\Phi(x_T)] + 4\alpha\kappa^2\delta'_0 + 7\alpha\kappa^2T\epsilon_\Delta + 4\alpha\kappa^2T\pi_\delta(d_1, d_2, \mu_1, \mu_2) + \alpha T\pi(d_1, d_2, \mu_1, \mu_2). \end{aligned} \quad (93)$$

Letting  $\alpha = \frac{1}{8L}$ , we obtain

$$\frac{\alpha}{2} - \frac{L\alpha^2}{2} - 12L^2\alpha^3 = \frac{1}{32L}. \quad (94)$$

Substituting eq. (93) into eq. (94) and applying Assumption 1 yield

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] & \leq 32L(\Phi(x_0) - \Phi^*) + 16\kappa^2\delta'_0 + 28\kappa^2T\epsilon_\Delta + 16\kappa^2T\pi_\delta(d_1, d_2, \mu_1, \mu_2) \\ & \quad + 4T\pi(d_1, d_2, \mu_1, \mu_2). \end{aligned} \quad (95)$$

Substituting eq. (95) and eq. (90) into eq. (86) yields

$$\begin{aligned} & \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\Phi(x_t)\|_2^2] \\ & \leq 6\kappa^2 \sum_{t=0}^{T-1} \delta'_t + 6T\epsilon_\Delta + 3 \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] + 3T\pi(d_1, d_2, \mu_1, \mu_2) \\ & \leq 12\kappa^2\delta'_0 + 21\kappa^2T\epsilon_\Delta + 4 \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|_2^2] + 12\kappa^2T\pi_\delta(d_1, d_2, \mu_1, \mu_2) + 3T\pi(d_1, d_2, \mu_1, \mu_2) \\ & \leq 128L(\Phi(x_0) - \Phi^*) + 76\kappa^2\delta'_0 + 133\kappa^2T\epsilon_\Delta + 76\kappa^2T\pi_\delta(d_1, d_2, \mu_1, \mu_2) + 19T\pi(d_1, d_2, \mu_1, \mu_2). \end{aligned} \quad (96)$$

Recall that  $L = (1 + \kappa)\ell$ . Then, eq. (96) implies

$$\begin{aligned} \mathbb{E}[\|\nabla\Phi(\hat{x})\|_2^2] & \leq 128(\kappa + 1)\ell \frac{\Phi(x_0) - \Phi^*}{T} + 133\kappa^2\epsilon_\Delta + \frac{76\kappa^2\delta'_0}{T} + 76\kappa^2\pi_\delta(d_1, d_2, \mu_1, \mu_2) \\ & \quad + 19\pi(d_1, d_2, \mu_1, \mu_2). \end{aligned}$$

If we let  $\delta'_0 \leq \frac{1}{\kappa}$ ,  $T = \max\{640(\kappa + 1)\ell \frac{\Phi(x_0) - \Phi^*}{\epsilon^2}, \frac{380\kappa}{\epsilon^2}\}$ , and let  $\mu_1$ ,  $\mu_2$  and  $\delta$  follow the same setting in Theorem 2, then we have

$$\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2] \leq \sqrt{\mathbb{E}[\|\nabla\Phi(\hat{x})\|_2^2]} \leq \epsilon.$$

Recall the sample complexity result of ZO-iSARSH in the finite-sum case in Appendix C.4. Then, we have  $T_0 = \mathcal{O}(d_2(\kappa + n) \log(\kappa))$ . The total sample complexity is given by

$$\begin{aligned} & T \cdot (S_{2,x} + S_{2,y}) \cdot m + \left\lceil \frac{T}{q} \right\rceil \cdot S_1 \cdot (d_1 + d_2) + T_0 \\ & \leq \Theta\left(\frac{\kappa}{\epsilon^2} \cdot (d_1 + d_2) \cdot \kappa\right) + \Theta\left(\left\lceil \frac{\kappa}{\epsilon^2} \right\rceil \cdot n \cdot (d_1 + d_2)\right) + \Theta(d_2(\kappa + n) \log(\kappa)) \\ & = \mathcal{O}\left((d_1 + d_2)(\kappa^2 + \kappa n)\epsilon^{-2}\right). \end{aligned}$$