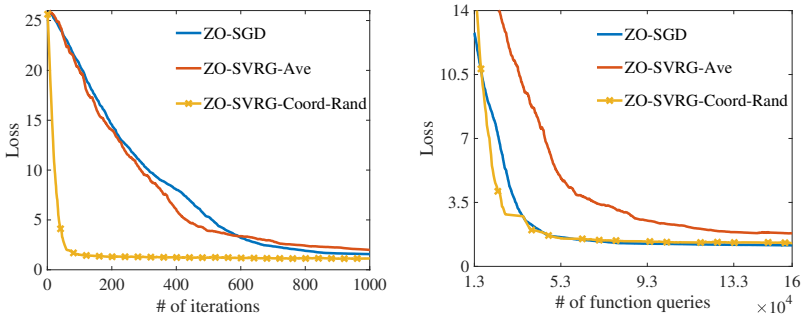
Supplementary Materials

A. Further Specification of Experiments and Additional Results

A.1. Generation of Black-Box Adversarial Examples

Parameter selection for algorithms under comparison. For ZO-SGD and ZO-SVRG-Ave, we adopt the implementations[3] from Liu et al. 2018b. As recommended by Liu et al. 2018b, we set the epoch length $q = 10$ for ZO-SVRG-Ave, and select the mini-batch size $|\mathcal{S}_2|$ from $\{5, 10, 50\}$ and the stepsize η from $\{1, 10, 20, 30, 40\}/d$ for both ZO-SGD and ZO-SVRG-Ave, and we present the best performance among these parameters, where $d = 28 \times 28$ is the input dimension. For SPIDER-SZO, we set the parameters by Theorem 8 in Fang et al. 2018. Namely, we choose the epoch length q from $\{30, 50, 80\}$, mini-batch size $|\mathcal{S}_2|$ from $\{5, 80, 700\}$, and η from $\{0.1, 0.01\}/\|\mathbf{v}^k\|$, and we present the best performance among these parameters. The parameters chosen for our ZO-SVRG-Coord-Rand, ZO-SVRG-Coord (based on our new analysis, which allows a larger stepsize with performance guarantee) and ZO-SPIDER-Coord are listed in Table 4. For all algorithms, we choose $|\mathcal{S}_1| = n$, and set the smoothing parameters $\beta = 0.01$ and $\delta = 0.001$.



(a) Loss vs iterations: $n = 10$ (b) Loss vs queries: $n = 10$

Figure 3. Comparison of different zeroth-order algorithms for generating black-box adversarial examples for digit "4" class

Table 3. Generated adversarial examples for digit "4" class, where image distortion is defined as $\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{a}_i^{adv} - \mathbf{a}_i\|^2$.

Image ID	4	6	19	24	27	33	42	48	49	56	Image distortion
ZO-SGD											11.46
Classified as	**9**	**8**	**1**	**3**	**2**	**2**	**9**	**9**	**9**	**9**	
ZO-SVRG-Ave											13.85
Classified as	**9**	**8**	**2**	**3**	**2**	**2**	**9**	**9**	**9**	**3**	
ZO-SVRG-Coord-Rand											11.21
Classified as	**9**	**8**	**2**	**3**	**2**	**2**	**9**	**9**	**9**	**9**	

A.2. Nonconvex logistic regression

Parameter selection for algorithms under comparison. For all algorithms, we choose fixed mini-batch sizes $|\mathcal{S}_1| = n$ and $|\mathcal{S}_2| = 128$, the epoch length $q = n/128$ for german dataset, and choose fixed mini-batch sizes $|\mathcal{S}_1| = 50 * 256$ and $|\mathcal{S}_2| = 256$, the epoch length $q = n/256$ for ijcnn1 dataset. In addition, we set the learning rate for all algorithms

[3] https://github.com/IBM/ZOSVRG-BlackBox-Adv

Table 4. Parameter settings for ZO-SVRG-Coord-Rand (left), ZO-SVRG-Coord (middle) and ZO-SPIDER-Coord (right).

Parameters	$n = 10$	$n = 100$
q	50	80
$\lvert\mathcal{S}_2\rvert$	80	700
η	0.102	0.663

Parameters	$n = 10$	$n = 100$
q	30	50
$\lvert\mathcal{S}_2\rvert$	5	70
η	0.102	0.255

Parameters	$n = 10$	$n = 100$
q	30	50
$\lvert\mathcal{S}_2\rvert$	5	70
η	0.064	0.255

according to their convergence guarantee. In specific, we choose $\eta = 0.8$ for ZO-SVRG-Coord-Rand, ZO-SPIDER-Coord, ZO-SVRG-Coord, and choose $\eta = 0.8/d$ for ZO-SGD, ZO-SVRG-Ave, and set $\eta = 0.8\sqrt{\epsilon}/\|v_k\|$ for SPIDER-SZO, as specified in Fang et al. 2018.

B. Zeroth-Order Nonconvex Nonsmooth Composite Optimization

Zeroth-order optimization has been studied for nonconvex and nonsmooth objective function in (Ghadimi et al., 2016), where a zeroth-order stochastic algorithm named RSPGF has been proposed. Here, we propose a zeroth-order stochastic variance-reduced algorithm for the same objective function, and show that it order-wisely outperforms RSPGF.

B.1. PROX-ZO-SPIDER-Coord for Composite Optimization

In this subsection, we extend our study of ZO-SPIDER-Coord to the following nonconvex and nonsmooth composite problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \Psi(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x}), f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \tag{15}$$

where each $f_i(\mathbf{x})$ is smooth and nonconvex, $h(\mathbf{x})$ is a nonsmooth convex function (e.g., $h(\mathbf{x}) = \lambda\|\mathbf{x}\|_1, \lambda > 0$). To address the nonsmooth term $h(\mathbf{x})$ in the objective function (15), we propose PROX-ZO-SPIDER-Coord algorithm, which replaces line 8 in Algorithm 2 by a proximal gradient step

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \big\{\langle \mathbf{v}^k, \mathbf{x}\rangle + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{x}^k\|^2 + h(\mathbf{x})\big\}.$$

Similarly to Ghadimi et al. 2016, we define

$$G(\mathbf{x}, \nabla f(\mathbf{x}), \eta) = \frac{1}{\eta}\left(\mathbf{x} - \mathbf{x}^+\right) \tag{16}$$

as a generalized projected gradient of $\Psi(\cdot)$ at the point $\mathbf{x}$ and use it to characterize the convergence criterion, where the point $\mathbf{x}^+$ is given by the proximal mapping

$$\mathbf{x}^+ = \arg\min_{\mathbf{z}\in\mathbb{R}^d} \left\{\langle \nabla f(\mathbf{x}), \mathbf{z}\rangle + \frac{1}{2\eta}\|\mathbf{z} - \mathbf{x}\|^2 + h(\mathbf{z})\right\}.$$

Based on the above notations, we provide the following convergence guarantee for PROX-ZO-SPIDER-Coord.

Theorem 5. *Let Assumption 1 hold, and we choose the same parameters as in Corollary 3. Then our PROX-ZO-SPIDER-Coord satisfies* $\mathbb{E}\|G(\mathbf{x}^\zeta, \nabla f(\mathbf{x}^\zeta), \eta)\|^2 \leq (60\Delta_\psi L + 80 + 69\sigma^2)/K + 138/K^2$, *where* $0 < \Delta_\psi = \psi(\mathbf{x}^0) - \psi(\mathbf{x}^*) < \infty$ *and* $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \psi(\mathbf{x})$.

To achieve $\mathbb{E}\|G(\mathbf{x}^\zeta, \nabla f(\mathbf{x}^\zeta), \eta)\|^2 \leq \epsilon$, *the number of function queries is at most* $\mathcal{O}\left(\min\left\{n^{1/2}d\epsilon^{-1}, d\epsilon^{-3/2}\right\}\right)$.

Let us compare our PROX-ZO-SPIDER-Coord algorithm with the randomized stochastic projected gradient free algorithm RSPGF, introduced by Ghadimi et al. 2016. Casting Corollary 8 in Ghadimi et al. 2016 to the setting of our Theorem 5 yields $\mathbb{E}\|G(\mathbf{x}^\zeta, \nabla f(\mathbf{x}^\zeta), \eta)\|^2 \leq \mathcal{O}\left(\frac{d}{\widetilde{K}} + \frac{\sqrt{d}}{\sqrt{\widetilde{K}}}\right)$, where $\widetilde{K}$ is the total number of function queries. Thus, RSPGF requires at most $\mathcal{O}(d/\epsilon^2)$ function queries to achieve $\mathbb{E}\|G(\mathbf{x}^\zeta, \nabla f(\mathbf{x}^\zeta), \eta)\|^2 \leq \epsilon < 1$. As a comparison, the function query complexity of PROX-ZO-SPIDER-Coord outperforms that of RSPGF (Ghadimi et al., 2016) by a factor of $\mathcal{O}(\max\{\epsilon^{-1}n^{-1/2}, \epsilon^{-1/2}\})$.

C. Zeroth-Order Variance-Reduced Algorithms for Convex Optimization

In this paper, we have proposed two new zeroth-order variance-reduced algorithms ZO-SVRG-Coord-Rand and ZO-SPIDER-Coord, and have studied their performance for nonconvex optimization. In this section, we study the performance of these two algorithms for convex optimization, where each individual function $f_i(\cdot)$ is convex. We note that there was no proven convergence guarantee for previously proposed zeroth-order SVRG-based and SPIDER-based algorithms for convex optimization.

C.1. ZO-SVRG-Coord-Rand-C Algorithm

In this subsection, we explore the convergence performance of ZO-SVRG-Coord-Rand for convex optimization. To fully utilize the convexity of the objective function, we propose a variant of our ZO-SVRG-Coord-Rand, which we refer to as ZO-SVRG-Coord-Rand-C. Differently from ZO-SVRG-Coord-Rand, the outer-loop iteration (i.e., $k \mod q = 0$) of ZO-SVRG-Coord-Rand-C chooses $\mathbf{x}^k$ from $\{\mathbf{x}^{k-q}, ..., \mathbf{x}^{k-1}\}$ uniformly at random, which is a typical treatment used in convex first-order optimization (Reddi et al., 2016a; Nguyen et al., 2017a). In the meanwhile, the inner-loop iteration of ZO-SVRG-Coord-Rand-C is the same as single-sample ZO-SVRG-Coord-Rand, which computes $\mathbf{v}^k = \hat{\nabla}_{\text{rand}} f_{i_k}(\mathbf{x}^k; \mathbf{u}^k) - \hat{\nabla}_{\text{rand}} f_{i_k}(\mathbf{x}^{qk_0}; \mathbf{u}^k) + \mathbf{v}^{qk_0}$ with a single sample i_k drawn from $[n]$ and a smoothing vector $\mathbf{u}^k$ drawn from the uniform distribution over the unit sphere. .

The following theorem provides the function query complexity for ZO-SVRG-Coord-Rand-C.

Theorem 6. *Under Assumption 1, let $\eta = 1/(27dL), \beta = \epsilon/(c_\beta dL), \delta = \epsilon/(c_\delta \sqrt{d}L), q = c_q d/\epsilon$, $h = \log_2(c_h/\epsilon))$ and $|\mathcal{S}| = \min\{n, \lceil c_s/\epsilon \rceil\}$, where $c_q, c_h, c_\beta, c_\delta$ and c_s are sufficiently large positive constants. Then, to achieve an ϵ-accuracy solution, i.e., $\mathbb{E}(f(\mathbf{x}^K) - f(\mathbf{x}^*)) \leq \epsilon$, the number of function queries required by ZO-SVRG-Coord-Rand-C algorithm is at most $\mathcal{O}(d \min\{n, 1/\epsilon\} \log(1/\epsilon))$.*

Let us compare our result with that of ZO-SGD given by Ghadimi & Lan 2013. Casting Corollary 3.3 in Ghadimi & Lan 2013 under the setting of our Corollary 6 implies that the function query complexity of ZO-SGD is $\mathcal{O}(d/\epsilon^2)$, which is worse than that of our ZO-SVRG-Coord-Rand-C by a factor of $\widetilde{\mathcal{O}}(\max\{\epsilon^{-2}n^{-1}, \epsilon^{-1}\})$.

C.2. ZO-SPIDER-Coord-C Algorithm

In this subsection, we generalize our ZO-SPIDER-Coord to solving convex optimization problem, and proposes the ZO-SPIDER-Coord-C algorithm. ZO-SPIDER-Coord-C has the same outer-loop iteration as ZO-SVRG-Coord-Rand-C, but updates $\mathbf{v}^k$ in a different way by $\mathbf{v}^k = \hat{\nabla}_{\text{coord}} f_{i_k}(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f_{i_k}(\mathbf{x}^{k-1}) + \mathbf{v}^{k-1}$ at each inner-loop iteration.

Based on Lemma 6, we obtain the following complexity result for ZO-SPIDER-Coord-C.

Theorem 7. *Under Assumption 1, let $\eta = 1/(24L), q = c_q/\epsilon$, $h = \log_2(c_h/\epsilon)), \delta = \epsilon/(c_q\sqrt{d}L)$ and $|\mathcal{S}| = \min\{n, \lceil c_s/\epsilon \rceil\}$, where c_q, c_h and c_s are sufficiently large positive constants. Then, to achieve an ϵ-accuracy solution, i.e., $\mathbb{E}\|\nabla f(\mathbf{x}^K)\|^2 \leq \epsilon$, the number of function queries required by ZO-SPIDER-Coord-C is at most $\mathcal{O}(d \min\{n, 1/\epsilon\} \log(1/\epsilon))$*

Note that ZO-SPIDER-Coord-C achieves the same function query complexity as that of ZO-SVRG-Coord-Rand-C, and improves that of ZO-SGD (Ghadimi & Lan, 2013) by a factor of $\widetilde{\mathcal{O}}(\max\{\epsilon^{-2}n^{-1}, \epsilon^{-1}\})$ w.r.t. stationary gap $\mathbb{E}\|\nabla f(\mathbf{x}^K)\|^2$. The detailed comparison among our algorithms and other exiting algorithms is summarized in Table 5.

Table 5. Comparison of zeroth-order algorithms in terms of the function query complexity for convex optimization.

Algorithms		Function query complexity	Function value convergence
ZO-SGD	(Ghadimi & Lan, 2013)	$\mathcal{O}\left(\frac{d}{\epsilon^2}\right)$	✓
ZSCG	(Balasubramanian & Ghadimi, 2018)	$\mathcal{O}\left(\frac{d}{\epsilon^3}\right)$	✓
M-ZSCG	(Balasubramanian & Ghadimi, 2018)	$\mathcal{O}\left(\frac{d}{\epsilon^2}\right)$	✓
ZO-SPIDER-Coord-C	(This work)	$\mathcal{O}(\min\{dn, \frac{d}{\epsilon}\} \log\left(\frac{1}{\epsilon}\right))$	✗
ZO-SVRG-Coord-Rand-C	(This work)	$\mathcal{O}(\min\{dn, \frac{d}{\epsilon}\} \log\left(\frac{1}{\epsilon}\right))$	✓

Technical Proofs

D. Proof for ZO-SVRG-Coord-Rand

D.1. Auxiliary Lemmas

Before proving our main results, we first establish three useful lemmas.

Lemma 3. *For any given smoothing parameter $\delta > 0$ and any $\mathbf{x} \in \mathbb{R}^d$, we have*

$$\|\hat{\nabla}_{\text{coord}} f(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq L^2 d\delta^2.$$

Proof. Applying the mean value theorem (MVT) to the gradient $\nabla f(\mathbf{x})$, we have, for any given $\delta > 0$,

$$\begin{aligned}
\|\hat{\nabla}_{\text{coord}} f(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 &= \Big\| \frac{1}{2\delta} \sum_{i=1}^{d} (2\delta \mathbf{e}_i \mathbf{e}_i^T \nabla f(\mathbf{x} + (2t_i - 1)\delta \mathbf{e}_i)) - \nabla f(\mathbf{x}) \Big\|^2, \text{ for } 0 < t_i < 1, \\
&\overset{\text{(i)}}{=} \sum_{i=1}^{d} \Big\| \mathbf{e}_i \mathbf{e}_i^T (\nabla f(\mathbf{x} + (2t_i - 1)\delta \mathbf{e}_i) - \nabla f(\mathbf{x})) \Big\|^2 \\
&\leq \sum_{i=1}^{d} \Big\| \nabla f(\mathbf{x} + (2t_i - 1)\delta \mathbf{e}_i) - \nabla f(\mathbf{x}) \Big\|^2 \overset{\text{(ii)}}{\leq} L^2 \sum_{i=1}^{d} \|(2t_i - 1)\delta \mathbf{e}^i\|^2 \leq L^2 d\delta^2
\end{aligned}$$

where (i) follows from the definition of $\mathbf{e}_i$ and Euclidean norm, and (ii) follows from Assumption 1. □

Lemma 4. *For any given $k_0 \leq \lfloor K/q \rfloor$, we have*

$$\mathbb{E}\|\mathbf{v}^{qk_0} - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0})\|^2 \leq \frac{3I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|} \left(2L^2 d\delta^2 + \sigma^2\right),$$

where $I(\cdot)$ is the indicator function.

Proof. To simplify notation, we let $\mathbf{z}_j = \hat{\nabla}_{\text{coord}} f_j(\mathbf{x}^{qk_0}) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0})$ and $I_j = I(j \in \mathcal{S}_1)$, where $I(\cdot)$ is the indicator function. First note that $\mathbb{E}(I_j^2) = \frac{|\mathcal{S}_1|}{n}$ and $\mathbb{E} I_i I_j = C_{|\mathcal{S}_1|}^2 / C_n^2 = \frac{|\mathcal{S}_1|(|\mathcal{S}_1|-1)}{n(n-1)}, i \neq j$. Then, based on the above equalities, we have

$$\begin{aligned}
\mathbb{E}\|\mathbf{v}^{qk_0} - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0})\|^2 &= \frac{1}{|\mathcal{S}_1|^2} \left(\sum_{j=1}^{n} \mathbb{E} I_j^2 \Big\| \mathbf{z}_j \Big\|^2 + \sum_{i \neq j} \mathbb{E} I_i I_j \langle \mathbf{z}_i, \mathbf{z}_j \rangle \right) \\
&= \frac{1}{|\mathcal{S}_1|^2} \left(\frac{|\mathcal{S}_1|}{n} \sum_{j=1}^{n} \Big\| \mathbf{z}_j \Big\|^2 + \frac{|\mathcal{S}_1|(|\mathcal{S}_1| - 1)}{n(n-1)} \sum_{i \neq j} \langle \mathbf{z}_i, \mathbf{z}_j \rangle \right) \\
&= \frac{1}{|\mathcal{S}_1|^2} \left(\left(\frac{|\mathcal{S}_1|}{n} - \frac{|\mathcal{S}_1|(|\mathcal{S}_1| - 1)}{n(n-1)} \right) \sum_{j=1}^{n} \Big\| \mathbf{z}_j \Big\|^2 + \frac{|\mathcal{S}_1|(|\mathcal{S}_1| - 1)}{n(n-1)} \Big\| \sum_{j=1}^{n} \mathbf{z}_j \Big\|^2 \right) \\
&= \frac{n - |\mathcal{S}_1|}{(n-1)|\mathcal{S}_1|} \frac{1}{n} \sum_{j=1}^{n} \Big\| \mathbf{z}_j \Big\|^2 + \frac{(|\mathcal{S}_1| - 1)}{n(n-1)|\mathcal{S}_1|} \Big\| \sum_{j=1}^{n} \mathbf{z}_j \Big\|^2 \\
&\leq \frac{I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|} \frac{1}{n} \sum_{j=1}^{n} \Big\| \mathbf{z}_j \Big\|^2 + \frac{1}{n^2} \Big\| \sum_{j=1}^{n} \mathbf{z}_j \Big\|^2 \\
&= \frac{I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|} \frac{1}{n} \sum_{j=1}^{n} \Big\| \hat{\nabla}_{\text{coord}} f_j(\mathbf{x}^{qk_0}) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0}) \Big\|^2 \\
&\leq \frac{3I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|} \frac{1}{n} \sum_{j=1}^{n} \Big(\|\hat{\nabla}_{\text{coord}} f_j(\mathbf{x}^{qk_0}) - \nabla f_j(\mathbf{x}^{qk_0})\|^2 + \|\nabla f_j(\mathbf{x}^{qk_0}) - \nabla f(\mathbf{x}^{qk_0})\|^2 \\
&\quad + \|\hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0}) - \nabla f(\mathbf{x}^{qk_0})\|^2 \Big) \leq \frac{3I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|} \left(2L^2 d\delta^2 + \sigma^2\right),
\end{aligned}$$

where the last inequality follows from Assumption 1 and Lemma 3. Then, the proof is complete. □

Lemma 5. *Let $f_\beta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim \mathrm{U}_B}\left(f(\mathbf{x}+\beta\mathbf{u})\right)$ be a smooth approximation of $f(\mathbf{x})$, where U_B is the uniform distribution over the d-dimensional unit Euclidean ball B. Then,*

(1) $|f_\beta(\mathbf{x}) - f(\mathbf{x})| \leq \frac{\beta^2 L}{2}$ *and* $\|\nabla f_\beta(\mathbf{x}) - \nabla f(\mathbf{x})\| \leq \frac{\beta L d}{2}$ *for any* $\mathbf{x}\in\mathbb{R}^d$

(2) $\mathbb{E}_k\left(\frac{1}{|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|}\hat{\nabla}_{\text{rand}} f_{a_j}(\mathbf{x};\mathbf{u}_j^k)\right) = \nabla f_\beta(\mathbf{x})$, *where* $\mathbf{x}$ *is either* $\mathbf{x}^k$ *or* $\mathbf{x}^{qk_0}$

(3) $\mathbb{E}_k\left\|\hat{\nabla}_{\text{rand}} f_{a_j}(\mathbf{x}^k;\mathbf{u}_j^k) - \hat{\nabla}_{\text{rand}} f_{a_j}(\mathbf{x}^{qk_0};\mathbf{u}_j^k)\right\|^2$
$\leq 3d^2\mathbb{E}_k\langle \nabla f_{a_j}(\mathbf{x}^k) - \nabla f_{a_j}(\mathbf{x}^{qk_0}), \mathbf{u}_j^k\rangle^2 + \frac{3L^2d^2\beta^2}{2}$
$\leq 3dL^2\|\mathbf{x}^k - \mathbf{x}^{qk_0}\|^2 + \frac{3L^2d^2\beta^2}{2}$,

where the shorthand $\mathbb{E}_k(\cdot) = \mathbb{E}(\cdot\,|\mathbf{x}^0,....,\mathbf{x}^k\,)$.

Proof. The proof of item (1) directly follows from Lemma 4.1 in Gao et al. 2014.

We next prove item (2). Based on the equation (3.4) in Gao et al. 2014, we have

$$\begin{aligned}\nabla f_\beta(\mathbf{x}^k) =& \mathbb{E}_{\mathbf{u}\sim \mathrm{U}_{S_p}}\left(\frac{d}{\beta} f(\mathbf{x}^k+\beta\mathbf{u})\mathbf{u}\right) \overset{\text{(i)}}{=} \mathbb{E}_{\mathbf{u}\sim \mathrm{U}_{S_p}}\left(\frac{d}{\beta}\Big(f(\mathbf{x}^k+\beta\mathbf{u}) - f(\mathbf{x}^k)\Big)\mathbf{u}\right)\\ =& \mathbb{E}_{\mathbf{u}\sim \mathrm{U}_{S_p}}\Big(\frac{1}{n}\sum_{i=1}^n \frac{d}{\beta}\left(f_i(\mathbf{x}^k+\beta\mathbf{u}) - f_i(\mathbf{x}^k)\right)\mathbf{u}\Big) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbf{u}\sim \mathrm{U}_{S_p}}\Big(\frac{d}{\beta}\left(f_i(\mathbf{x}^k+\beta\mathbf{u}) - f_i(\mathbf{x}^k)\right)\mathbf{u}\Big), \end{aligned} \tag{17}$$

where the random vector $\mathbf{u}$ is independent of $\mathbf{x}^k$, U_{S_p} is the uniform distribution over the unit sphere S_p and (i) follows from the fact that $\mathbb{E}_{\mathbf{u}\sim \mathrm{U}_{S_p}}\left(f(\mathbf{x}^k)\mathbf{u}\right) = 0$. To simplify notation, we let $\mathbb{E}_k(\cdot) = \mathbb{E}(\cdot|\mathbf{x}^1,...,\mathbf{x}^k)$. Conditioned on $\mathbf{x}^0,....,\mathbf{x}^k$ and noting that the random samples in $\mathcal{S}_2$ and $\mathbf{u}_j^k, j=1,...,|\mathcal{S}_2|$ generated at the k^{th} iteration are independent of $\mathbf{x}^0,....,\mathbf{x}^k$, we have

$$\begin{aligned}&\mathbb{E}_k\left(\frac{d}{\beta|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|}(f_{a_j}(\mathbf{x}^k+\beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}^k))\mathbf{u}_j^k\right) = \frac{1}{|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|}\mathbb{E}_k\left(\frac{d}{\beta}(f_{a_j}(\mathbf{x}^k+\beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}^k))\mathbf{u}_j^k\right)\\ &= \frac{1}{|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|}\mathbb{E}_k\left(\mathbb{E}_{a_j}\left(\frac{d}{\beta}(f_{a_j}(\mathbf{x}^k+\beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}^k))\mathbf{u}_j^k\,\Big|\,\mathbf{u}_j^k\right)\right)\\ &\overset{\text{(i)}}{=} \frac{1}{|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|}\mathbb{E}_k\left(\frac{1}{n}\sum_{i=1}^n\frac{d}{\beta}(f_i(\mathbf{x}^k+\beta\mathbf{u}_j^k) - f_i(\mathbf{x}^k))\mathbf{u}_j^k\right)\\ &\overset{\text{(ii)}}{=} \frac{1}{|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|}\left(\nabla f_\beta(\mathbf{x}^k)\right) = \nabla f_\beta(\mathbf{x}^k),\end{aligned} \tag{18}$$

where (i) follows from the definition of the set $\mathcal{S}_1$ and (ii) follows from (17). Taking steps similar to (18) and conditioning on $\mathbf{x}^0,...,\mathbf{x}^k$, we have

$$\mathbb{E}_k\left(\frac{d}{\beta|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|}(f_{a_j}(\mathbf{x}^{qk_0}+\beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}^k))\mathbf{u}_j^k\right) = \nabla f_\beta(\mathbf{x}^{qk_0}). \tag{19}$$

Our final step is to prove item (3). Note that

$$\begin{aligned}&\mathbb{E}\left\|\frac{d(f_{a_j}(\mathbf{x}^k+\beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}^k))}{\beta}\mathbf{u}_j^k - \frac{d(f_{a_j}(\mathbf{x}^{qk_0}+\beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}^{qk_0}))}{\beta}\mathbf{u}_j^k\right\|^2\\ &= d^2\mathbb{E}_k\left\|\frac{(f_{a_j}(\mathbf{x}^k+\beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}^k) - \langle\nabla f_{a_j}(\mathbf{x}^k),\beta\mathbf{u}_j^k\rangle)\mathbf{u}_j^k}{\beta} + \left(\langle\nabla f_{a_j}(\mathbf{x}^k),\mathbf{u}_j^k\rangle\mathbf{u}_j^k - \langle\nabla f_{a_j}(\mathbf{x}^{qk_0}),\mathbf{u}_j^k\rangle\mathbf{u}_j^k\right)\right.\\ &\quad\left. - \frac{(f_{a_j}(\mathbf{x}^{qk_0}+\beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}^{qk_0}) - \langle\nabla f_{a_j}(\mathbf{x}^{qk_0}),\beta\mathbf{u}_j^k\rangle)\mathbf{u}_j^k}{\beta}\right\|^2\end{aligned} \tag{20}$$

Then, using the inequality that $f_{a_j}(\mathbf{y}) - f_{a_j}(\mathbf{x}) - \langle \nabla f_{a_j}(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle \leq \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2$ in (20) yields

$$\begin{aligned}
&\mathbb{E}\left\|\frac{d(f_{a_j}(\mathbf{x}^k + \beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}^k))}{\beta}\mathbf{u}_j^k - \frac{d(f_{a_j}(\mathbf{x}^{qk_0} + \beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}^{qk_0}))}{\beta}\mathbf{u}_j^k\right\|^2 \\
\leq& 3d^2\mathbb{E}_k\|\langle \nabla f_{a_j}(\mathbf{x}^k), \mathbf{u}_j^k\rangle\mathbf{u}_j^k - \langle \nabla f_{a_j}(\mathbf{x}^{qk_0}), \mathbf{u}_j^k\rangle\mathbf{u}_j^k\|^2 + \frac{3L^2d^2\beta^2}{2} \\
\overset{\text{(i)}}{=}& 3d^2\mathbb{E}_k\langle \nabla f_{a_j}(\mathbf{x}^k) - \nabla f_{a_j}(\mathbf{x}^{qk_0}), \mathbf{u}_j^k\rangle^2 + \frac{3L^2d^2\beta^2}{2} \\
=& 3d^2\left(\nabla f_{a_j}(\mathbf{x}^k) - \nabla f_{a_j}(\mathbf{x}^{qk_0})\right)^T\mathbb{E}(\mathbf{u}_j^k(\mathbf{u}_j^k)^T)\left(\nabla f_{a_j}(\mathbf{x}^k) - \nabla f_{a_j}(\mathbf{x}^{qk_0})\right) + \frac{3L^2d^2\beta^2}{2}
\end{aligned} \tag{21}$$

where (i) follows from the fact that $\|\mathbf{u}_j^k\| = 1$. Based on the definition of $\mathbf{u}_j^k$, we rewrite $\mathbf{u}_j^k = \mathbf{r}/\|\mathbf{r}\|$ and define a matrix $\mathbf{U} = \mathbb{E}(\mathbf{u}_j^k(\mathbf{u}_j^k)^T)$, where $\mathbf{r}$ is a d-dimensional Gaussian standard random vector. Let $\mathbf{r}(i)$ denote the i^{th} entry of $\mathbf{r}$, and $\mathbf{U}(i,j)$ denote $(i,j)^{th}$ entry of $\mathbf{U}$. Then, we have, for $i = 1, ..., d$

$$\mathbf{U}(i,i) = \int \frac{\mathbf{r}(i)^2}{\sum_{t=1}^d \mathbf{r}(t)^2} e^{-\frac{\sum_{t=1}^d \mathbf{r}(t)^2}{2}} d\mathbf{r}(1)\cdots d.\mathbf{r}(d) \tag{22}$$

Since $\mathbf{r}(i), i = 1, ..., d$ are i.i.d. standard Gaussian random variables, we have

$$\mathbf{U}(1,1) = \cdots \mathbf{U}(d,d), \text{ and } \sum_{i=1}^d \mathbf{U}(i,i) = \int e^{-\frac{\sum_{t=1}^d \mathbf{r}(t)^2}{2}} d\mathbf{r}(1)\cdots d\mathbf{r}(d) = 1,$$

which implies that $\mathbf{U}(i,i) = 1/d$ for all $i = 1, ..., d$. In addition, for any $i \neq j$, we have

$$\mathbf{U}(i,j) = \int \frac{\mathbf{r}(i)\mathbf{r}(j)}{\sum_{t=1}^d \mathbf{r}(t)^2} e^{-\frac{\sum_{t=1}^d \mathbf{r}(t)^2}{2}} d\mathbf{r}(1)\cdots d\mathbf{r}(d),$$

which, noting the symmetry between $\mathbf{r}(i)$ and $\mathbf{r}(j)$, implies that $\mathbf{U}(i,j) = 0$. Combining the above two results yields that $\mathbf{U} = \frac{1}{d}\mathbf{I}_d$, where $\mathbf{I}_d$ is a d-dimensional identity matrix. Thus, plugging $\mathbb{E}(\mathbf{u}_j^k(\mathbf{u}_j^k)^T) = \mathbf{U} = \frac{1}{d}\mathbf{I}_d$ in (21) yields

$$\begin{aligned}
&\mathbb{E}\left\|\frac{d(f_{a_j}(\mathbf{x}^k + \beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}^k))}{\beta}\mathbf{u}_j^k - \frac{d(f_{a_j}(\mathbf{x}^{qk_0} + \beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}^{qk_0}))}{\beta}\mathbf{u}_j^k\right\|^2 \\
\leq& 3d\|\nabla f_{a_j}(\mathbf{x}^k) - \nabla f_{a_j}(\mathbf{x}^{qk_0})\|^2 + \frac{3L^2d^2\beta^2}{2} \leq 3dL^2\|\mathbf{x}^k - \mathbf{x}^{qk_0}\|^2 + \frac{3L^2d^2\beta^2}{2},
\end{aligned} \tag{23}$$

which finishes the proof. □

D.2. Proof of Lemma 1

Using Lemmas 3, 4, 5, we now prove Lemma 1. Based on the updating step of Algorithm 1, we obtain

$$\begin{aligned}
\mathbf{v}^k - \nabla f_\beta(\mathbf{x}^k) =& \frac{d}{\beta|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|}(f_{a_j}(\mathbf{x}^k + \beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}^k))\mathbf{u}_j^k - \nabla f_\beta(\mathbf{x}^k) \\
&- \frac{d}{\beta|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|}(f_{a_j}(\mathbf{x}^{qk_0} + \beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}^{qk_0}))\mathbf{u}_j^k + \nabla f_\beta(\mathbf{x}^{qk_0}) + \mathbf{v}^{qk_0} - \nabla f_\beta(\mathbf{x}^{qk_0}).
\end{aligned} \tag{24}$$

To simplify notation, we define

$$H_j(\mathbf{x}) = \frac{d}{\beta}(f_{a_j}(\mathbf{x} + \beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}))\mathbf{u}_j^k, \ \mathbf{x} = \mathbf{x}^k \text{ or } \mathbf{x}^{qk_0},$$

and use the shorthand $\mathbb{E}_k(\cdot)$ to denote $\mathbb{E}(\cdot \mid \mathbf{x}^0, ..., \mathbf{x}^k)$. Then, using (24), we obtain

$$\begin{aligned}
\mathbb{E}_k\|\mathbf{v}^k - \nabla f_\beta(\mathbf{x}^k)\|^2 \leq & \frac{2}{|\mathcal{S}_2|}\mathbb{E}_k\|H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0}) - (\nabla f_\beta(\mathbf{x}^k) - \nabla f_\beta(\mathbf{x}^{qk_0}))\|^2 \\
& + 2\sum_{i\neq j}\mathbb{E}_k\langle H_i(\mathbf{x}^k) - H_i(\mathbf{x}^{qk_0}) - (\nabla f_\beta(\mathbf{x}^k) - \nabla f_\beta(\mathbf{x}^{qk_0})), H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0}) \\
& - (\nabla f_\beta(\mathbf{x}^k) - \nabla f_\beta(\mathbf{x}^{qk_0}))\rangle + 2\|\mathbf{v}^{qk_0} - \nabla f_\beta(\mathbf{x}^{qk_0})\|^2 \\
\overset{\text{(i)}}{=} & \frac{2}{|\mathcal{S}_2|}\mathbb{E}_k\|H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0}) - (\nabla f_\beta(\mathbf{x}^k) - \nabla f_\beta(\mathbf{x}^{qk_0}))\|^2 \\
& + 2\|\mathbf{v}^{qk_0} - \nabla f_\beta(\mathbf{x}^{qk_0})\|^2
\end{aligned} \tag{25}$$

where (i) follows from the fact that a_i and $\mathbf{u}_i^k$ are independent of a_j and $\mathbf{u}_j^k$ for any $i \neq j$, and from the following equalities

$$\begin{aligned}
\mathbb{E}_k(H_j(\mathbf{x}^k)) &= \mathbb{E}_{\mathbf{u}_j^k}\Big(\frac{d}{\beta}f(\mathbf{x}^k + \beta\mathbf{u}_j^k)\mathbf{u}_j^k\Big) = \nabla f_\beta(\mathbf{x}^k) \\
\mathbb{E}_k(H_j(\mathbf{x}^{qk_0})) &= \mathbb{E}_{\mathbf{u}_j^k}\Big(\frac{d}{\beta}f(\mathbf{x}^{qk_0} + \beta\mathbf{u}_j^k)\mathbf{u}_j^k\Big) = \nabla f_\beta(\mathbf{x}^{qk_0}).
\end{aligned} \tag{26}$$

Then, we further simplify (25) to obtain

$$\begin{aligned}
\mathbb{E}_k\|\mathbf{v}^k - \nabla f_\beta(\mathbf{x}^k)\|^2 \leq & \frac{2}{|\mathcal{S}_2|}\left(\mathbb{E}_k\|H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0})\|^2 + \mathbb{E}_k\|\nabla f_\beta(\mathbf{x}^k) - \nabla f_\beta(\mathbf{x}^{qk_0})\|^2\right) \\
& - \frac{4}{|\mathcal{S}_2|}\mathbb{E}_k\left\langle H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0}), \nabla f_\beta(\mathbf{x}^k) - \nabla f_\beta(\mathbf{x}^{qk_0})\right\rangle \\
& + 2\|\mathbf{v}^{qk_0} - \nabla f_\beta(\mathbf{x}^{qk_0})\|^2 \\
\overset{\text{(i)}}{\leq} & \frac{2}{|\mathcal{S}_2|}\mathbb{E}_k\|H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0})\|^2 + 2\|\mathbf{v}^{qk_0} - \nabla f_\beta(\mathbf{x}^{qk_0})\|^2 \\
\leq & \frac{2}{|\mathcal{S}_2|}\mathbb{E}_k\|H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0})\|^2 + 6\|\mathbf{v}^{qk_0} - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0})\|^2 \\
& + 6\|\nabla f_\beta(\mathbf{x}^{qk_0}) - \nabla f(\mathbf{x}^{qk_0})\|^2 + 6\|\nabla f(\mathbf{x}^{qk_0}) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0})\|^2 \\
\overset{\text{(ii)}}{\leq} & \frac{2}{|\mathcal{S}_2|}\mathbb{E}_k\|H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0})\|^2 + \frac{18I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2d\delta^2 + \sigma^2\right) \\
& + 6L^2d\delta^2 + \frac{3\beta^2L^2d^2}{2}
\end{aligned} \tag{27}$$

where (i) follows from (26) and (ii) follows from Lemmas 5, 3 and 4. Then, based on item (3) in Lemma 5, we obtain

$$\begin{aligned}
\mathbb{E}_k\|H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0})\|^2 &= \mathbb{E}_k\left\|\frac{d(f_{a_j}(\mathbf{x}^k + \beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}^k))}{\beta}\mathbf{u}_j^k - \frac{d(f_{a_j}(\mathbf{x}^{qk_0} + \beta\mathbf{u}_j^k) - f_{a_j}(\mathbf{x}^{qk_0}))}{\beta}\mathbf{u}_j^k\right\|^2 \\
&\leq 3dL^2\|\mathbf{x}^k - \mathbf{x}^{qk_0}\|^2 + \frac{3L^2d^2\beta^2}{2}.
\end{aligned} \tag{28}$$

Combining (27) and (28) finishes the proof.

D.3. Proof of Theorem 1

Since $K = qh$ and $\nabla f_\beta(\mathbf{x})$ is L-Lipschitz, we have, for $qm \leq k \leq q(m+1) - 1, m = 0, ..., h-1$

$$f_\beta(\mathbf{x}^{k+1}) \leq f_\beta(\mathbf{x}^k) + \langle\nabla f_\beta(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k\rangle + \frac{L\eta^2}{2}\|\mathbf{v}^k\|^2 = f_\beta(\mathbf{x}^k) - \eta\langle\nabla f_\beta(\mathbf{x}^k), \mathbf{v}^k\rangle + \frac{L\eta^2}{2}\|\mathbf{v}^k\|^2.$$

Taking the expectation over the above inequality and noting from Lemma 5 that $\mathbb{E}(\mathbf{v}^k \mid \mathbf{x}^0, ..., \mathbf{x}^k) = \nabla f_\beta(\mathbf{x}^k) - \nabla f_\beta(\mathbf{x}^{qm}) + \mathbf{v}^{qm}$, we have

$$\begin{aligned}
\mathbb{E}f_\beta(\mathbf{x}^{k+1}) \leq & \mathbb{E}f_\beta(\mathbf{x}^k) - \eta\mathbb{E}\langle \nabla f_\beta(\mathbf{x}^k) - \nabla f_\beta(\mathbf{x}^{qm}) + \mathbf{v}^{qm}, \nabla f_\beta(\mathbf{x}^k)\rangle + \frac{L\eta^2}{2}\mathbb{E}\|\mathbf{v}^k\|^2 \\
\leq & \mathbb{E}f_\beta(\mathbf{x}^k) - \eta\mathbb{E}\|\nabla f_\beta(\mathbf{x}^k)\|^2 + \eta\mathbb{E}\langle \nabla f_\beta(\mathbf{x}^{qm}) - \mathbf{v}^{qm}, \nabla f_\beta(\mathbf{x}^k)\rangle + \frac{L\eta^2}{2}\mathbb{E}\|\mathbf{v}^k\|^2 \\
\leq & \mathbb{E}f_\beta(\mathbf{x}^k) - \eta\mathbb{E}\|\nabla f_\beta(\mathbf{x}^k)\|^2 + \frac{\eta}{2}\mathbb{E}\|\nabla f_\beta(\mathbf{x}^{qm}) - \mathbf{v}^{qm}\|^2 + \frac{\eta}{2}\mathbb{E}\|\nabla f_\beta(\mathbf{x}^k)\|^2 + \frac{L\eta^2}{2}\mathbb{E}\|\mathbf{v}^k\|^2 \\
\leq & \mathbb{E}f_\beta(\mathbf{x}^k) - \frac{\eta}{2}\mathbb{E}\|\nabla f_\beta(\mathbf{x}^k)\|^2 + \frac{\eta}{2}\mathbb{E}\|\nabla f_\beta(\mathbf{x}^{qm}) - \mathbf{v}^{qm}\|^2 + \frac{L\eta^2}{2}\mathbb{E}\|\mathbf{v}^k\|^2 \\
\overset{\text{(i)}}{\leq} & \mathbb{E}f_\beta(\mathbf{x}^k) - \frac{\eta}{2}\left(\frac{1}{2}\|\nabla f(\mathbf{x}^k)\|^2 - \frac{\beta^2 L^2 d^2}{4}\right) + \frac{\eta}{2}\mathbb{E}\|\nabla f_\beta(\mathbf{x}^{qm}) - \mathbf{v}^{qm}\|^2 + \frac{L\eta^2}{2}\mathbb{E}\|\mathbf{v}^k\|^2
\end{aligned} \tag{29}$$

where (i) follows from the inequality that $\|\mathbf{a}\|^2 \geq \frac{1}{2}\|\mathbf{b}\|^2 - \|\mathbf{b}-\mathbf{a}\|^2$. Using an approach similar to (27), we obtain

$$\begin{aligned}
\mathbb{E}\|\nabla f_\beta(\mathbf{x}^{qm}) - \mathbf{v}^{qm}\|^2 \leq & 3\|\mathbf{v}^{qm} - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qm})\|^2 + 3\|\nabla f_\beta(\mathbf{x}^{qm}) - \nabla f(\mathbf{x}^{qm})\|^2 \\
& + 3\|\nabla f(\mathbf{x}^{qm}) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qm})\|^2 \\
\leq & \frac{9I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right) + 3L^2 d\delta^2 + \frac{3\beta^2 L^2 d^2}{4},
\end{aligned} \tag{30}$$

which, in conjunction with (29), implies that

$$\begin{aligned}
\mathbb{E}f_\beta(\mathbf{x}^{k+1}) \leq & \mathbb{E}f_\beta(\mathbf{x}^k) - \frac{\eta}{4}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \frac{\eta}{2}\left(\beta^2 L^2 d^2 + \frac{9I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right) + 3L^2 d\delta^2\right) \\
& + \frac{3L\eta^2}{2}\mathbb{E}\left(\|\nabla f_\beta(\mathbf{x}^k) - \mathbf{v}^k\|^2 + \|\nabla f_\beta(\mathbf{x}^k) - \nabla f(\mathbf{x}^k)\|^2 + \|\nabla f(\mathbf{x}^k)\|^2\right) \\
\overset{\text{(i)}}{\leq} & \mathbb{E}f_\beta(\mathbf{x}^k) - \left(\frac{\eta}{4} - \frac{3L\eta^2}{2}\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \frac{\eta}{2}\left(\beta^2 L^2 d^2 + \frac{9I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right) + 3L^2 d\delta^2\right) \\
& + \frac{3\eta^2 L^3 d^2 \beta^2}{8} + \frac{3L\eta^2}{|\mathcal{S}_2|}\left(3dL^2\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{qm}\|^2 + \frac{3L^2\beta^2 d^2}{2}\right) \\
& + \frac{3L\eta^2}{2}\left(\frac{18I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right) + 6L^2 d\delta^2 + \frac{3\beta^2 L^2 d^2}{2}\right)
\end{aligned} \tag{31}$$

where (i) follows from Lemma 1. To simplify notation, we define

$$\chi = \beta^2 L^2 d^2 + \frac{9I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right) + 3L^2 d\delta^2 \tag{32}$$

which, in conjunction with (31), implies that

$$\begin{aligned}
\mathbb{E}f_\beta(\mathbf{x}^{k+1}) \leq & \mathbb{E}f_\beta(\mathbf{x}^k) - \left(\frac{\eta}{4} - \frac{3L\eta^2}{2}\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \frac{9dL^3\eta^2}{|\mathcal{S}_2|}\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{qm}\|^2 + \frac{3\eta^2 L}{2}\left(\frac{L^2 d^2 \beta^2}{4} + \frac{3L^2 d^2 \beta^2}{|\mathcal{S}_2|}\right) \\
& + \left(\frac{\eta}{2} + 3L\eta^2\right)\chi.
\end{aligned} \tag{33}$$

We introduce a Lyapunov function $R_k^m = \mathbb{E}\left(f_\beta(\mathbf{x}^k) + c_k^m\|\mathbf{x}^k - \mathbf{x}^{qm}\|^2\right)$ for $qm \leq k \leq q(m+1), m = 0, ..., h-1$,

where $\{c_k^m\}$ are constants such that $c_{q(m+1)}^m = 0$. Then, we obtain that for any $qm \leq k \leq q(m+1)-1$

$$\begin{aligned}
R_{k+1}^m =& \mathbb{E}\left(f_\beta(\mathbf{x}^{k+1}) + c_{k+1}^m\|\mathbf{x}^{k+1}-\mathbf{x}^k+\mathbf{x}^k-\mathbf{x}^{qm}\|^2\right)\\
\leq& \mathbb{E}f_\beta(\mathbf{x}^{k+1}) + c_{k+1}^m\eta^2\mathbb{E}\|\mathbf{v}^k\|^2 + c_{k+1}^m\mathbb{E}\|\mathbf{x}^k-\mathbf{x}^{qm}\|^2 - 2c_{k+1}^m\eta\mathbb{E}\langle\mathbf{v}^k,\mathbf{x}^k-\mathbf{x}^{qm}\rangle\\
\overset{\text{(i)}}{=}& \mathbb{E}f_\beta(\mathbf{x}^{k+1}) + c_{k+1}^m\eta^2\mathbb{E}\|\mathbf{v}^k\|^2 + c_{k+1}^m\mathbb{E}\|\mathbf{x}^k-\mathbf{x}^{qm}\|^2 - 2c_{k+1}^m\eta\mathbb{E}\langle\nabla f_\beta(\mathbf{x}^k)-\nabla f_\beta(\mathbf{x}^{qm})+\mathbf{v}^{qm},\mathbf{x}^k-\mathbf{x}^{qm}\rangle\\
\overset{\text{(ii)}}{\leq}& \mathbb{E}f_\beta(\mathbf{x}^{k+1}) + c_{k+1}^m\eta^2\mathbb{E}\|\mathbf{v}^k\|^2 + (c_{k+1}^m + c_{k+1}^m\eta g)\mathbb{E}\|\mathbf{x}^k-\mathbf{x}^{qm}\|^2\\
&+\frac{2c_{k+1}^m\eta}{g}\mathbb{E}\left(\|\nabla f_\beta(\mathbf{x}^k)\|^2+\|\nabla f_\beta(\mathbf{x}^{qm})-\mathbf{v}^{qm}\|^2\right)\\
\leq& \mathbb{E}f_\beta(\mathbf{x}^{k+1}) + c_{k+1}^m\eta^2\mathbb{E}\left(2\|\mathbf{v}^k-\nabla f_\beta(\mathbf{x}^k)\|^2+\frac{\beta^2L^2d^2}{2}\right) + \left(c_{k+1}^m + c_{k+1}^m\eta g\right)\mathbb{E}\|\mathbf{x}^k-\mathbf{x}^{qm}\|^2\\
&+\frac{4c_{k+1}^m\eta}{g}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2+\frac{c_{k+1}^m\eta\beta^2L^2d^2}{g}+\frac{2c_{k+1}^m\eta}{g}\chi\\
\overset{\text{(iii)}}{\leq}& \mathbb{E}f_\beta(\mathbf{x}^{k+1}) + \left(c_{k+1}^m + c_{k+1}^m\eta g+\frac{12c_{k+1}^m\eta^2L^2d}{|\mathcal{S}_2|}\right)\mathbb{E}\|\mathbf{x}^k-\mathbf{x}^{qm}\|^2 + c_{k+1}^m\eta^2\left(4\chi+\frac{6L^2\beta^2d^2}{|\mathcal{S}_2|}\right)\\
&+\frac{4c_{k+1}^m\eta}{g}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2+\frac{c_{k+1}^m\eta\beta^2L^2d^2}{g}+\frac{2c_{k+1}^m\eta}{g}\chi
\end{aligned}\tag{34}$$

where (i) follows from the fact that $\mathbb{E}(\mathbf{v}^k\,|\,\mathbf{x}^0,....,\mathbf{x}^k) = \nabla f_\beta(\mathbf{x}^k) - \nabla f_\beta(\mathbf{x}^{qm}) + \mathbf{v}^{qm}$, (ii) follows from the fact that $-2\langle\mathbf{a},\mathbf{b}\rangle \leq \|\mathbf{a}\|^2/g + g\|\mathbf{b}\|^2$ holds for any constant $g>0$ and (iii) follows from Lemma 1. Combining (33) and (34), we obtain that

$$\begin{aligned}
R_{k+1}^m \leq& \mathbb{E}f_\beta(\mathbf{x}^k) - \left(\frac{\eta}{4}-\frac{4c_{k+1}^m\eta}{g}-\frac{3L\eta^2}{2}\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2\\
&+\left(c_{k+1}^m + c_{k+1}^m\eta g+\frac{12c_{k+1}^m\eta^2L^2d}{|\mathcal{S}_2|}+\frac{9dL^3\eta^2}{|\mathcal{S}_2|}\right)\mathbb{E}\|\mathbf{x}^k-\mathbf{x}^{qm}\|^2\\
&+\left(\frac{\eta}{2}+\frac{2c_{k+1}^m\eta}{g}+4c_{k+1}^m\eta^2+3L\eta^2\right)\chi+\left(\frac{3\eta^2L}{2}\left(\frac{1}{4}+\frac{3}{|\mathcal{S}_2|}\right)+\frac{c_{k+1}^m\eta}{g}\right)L^2d^2\beta^2.
\end{aligned}\tag{35}$$

We define the following recursion for $qm \leq k \leq q(m+1)-1, m = 0, ..., h-1$

$$c_k^m = c_{k+1}^m + c_{k+1}^m\eta g + \frac{12c_{k+1}^m\eta^2L^2d}{|\mathcal{S}_2|} + \frac{9dL^3\eta^2}{|\mathcal{S}_2|}\tag{36}$$

which, in conjunction with (35), implies that

$$\begin{aligned}
R_{k+1}^m \leq& R_k^m - \left(\frac{\eta}{4}-\frac{4c_{k+1}^m\eta}{g}-\frac{3L\eta^2}{2}\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2\\
&+\left(\frac{\eta}{2}+\frac{2c_{k+1}^m\eta}{g}+4c_{k+1}^m\eta^2+3L\eta^2\right)\chi+\left(\frac{3\eta^2L}{2}\left(\frac{1}{4}+\frac{3}{|\mathcal{S}_2|}\right)+\frac{c_{k+1}^m\eta}{g}\right)L^2d^2\beta^2\\
\leq& R_k^m - \left(\frac{\eta}{4}-\frac{4c_{k+1}^m\eta}{g}-\frac{3L\eta^2}{2}\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2\\
&+\left(\frac{\eta}{2}+\frac{2c_{k+1}^m\eta}{g}+4c_{k+1}^m\eta^2+3L\eta^2\right)\chi+\left(6\eta^2L+\frac{c_{k+1}^m\eta}{g}\right)L^2d^2\beta^2
\end{aligned}\tag{37}$$

where the last inequality follows from the fact that $1/4 + 3/|\mathcal{S}_2| \leq 4$. Letting $\theta = \eta g + 12\eta^2dL^2/|\mathcal{S}_2|$ and noting that $c_{q(m+1)}^m = 0$, we obtain from (36) that for $qm \leq k \leq q(m+1)-1, m = 0, ..., h-1$

$$c_k^m \leq c_{qm}^m = \frac{9dL^3\eta^2}{|\mathcal{S}_2|}\frac{(1+\theta)^q-1}{\theta},$$

which, in conjunction with (37) and the parameter selection in (3), implies that

$$R_{k+1}^m \leq R_k^m - \lambda \mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \tau.$$

Telescoping the above inequality over k from qm to $q(m+1)-1$ and noting that $R_{qm}^m = \mathbb{E} f_\beta(\mathbf{x}^{qm})$ and $R_{q(m+1)}^m = \mathbb{E} f_\beta(\mathbf{x}^{q(m+1)})$, we obtain

$$\mathbb{E} f_\beta(\mathbf{x}^{q(m+1)}) \leq \mathbb{E} f_\beta(\mathbf{x}^{qm}) - \lambda \sum_{k=qm}^{q(m+1)-1} \mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + q\tau.$$

Then, telescoping the above inequality over m from 0 to $h-1$, we obtain

$$\mathbb{E} f_\beta(\mathbf{x}^K) \leq \mathbb{E} f_\beta(\mathbf{x}^0) - \lambda \sum_{k=0}^{K} \mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + K\tau,$$

which can be rewritten as

$$\frac{1}{K+1}\sum_{k=0}^{K} \mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \leq \frac{f_\beta(\mathbf{x}^0) - f_\beta(\mathbf{x}_\beta^*)}{\lambda(K+1)} + \frac{\tau}{\lambda}, \tag{38}$$

where $\mathbf{x}_\beta^* = \arg\min_{\mathbf{x}} f_\beta(\mathbf{x})$. Since the output $\mathbf{x}^\zeta$ of Algorithm 1 is generated from $\{\mathbf{x}^0, ..., \mathbf{x}^K\}$ uniformly at random, we have

$$\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 = \frac{1}{K+1}\sum_{k=0}^{K} \|\nabla f(\mathbf{x}^k)\|^2,$$

which, in conjunction with (38), finishes the proof.

D.4. Proof of Corollary 1

We prove two cases with $n \leq K$ and $n > K$, separately.

First we suppose $n \leq K$. In this case, we have $|\mathcal{S}_1| = n$. Recall from (4) that

$$c = \frac{9dL^3\eta^2}{|\mathcal{S}_2|}\frac{(1+\theta)^q - 1}{\theta} \tag{39}$$

where $\theta = \eta g + 12\eta^2 dL^2/|\mathcal{S}_2|$. Based on the parameter selection in (6), we have

$$\theta \overset{\text{(i)}}{\leq} \frac{1}{2q} + \frac{3}{100}\frac{1}{q}\frac{1}{q} < \frac{1}{q} \text{ and } \theta > \frac{1}{2q}. \tag{40}$$

which, in conjunction with (39), yields

$$c \leq \frac{18(e-1)dL^3\eta^2 q}{|\mathcal{S}_2|} \leq \frac{9(e-1)L}{200q} \tag{41}$$

where (i) follows from the fact that $(1+\theta)^q \leq (1+1/q)^q < e$ and e is the Euler's number. Since $g = 4000d\eta^2 L^3 q/|\mathcal{S}_2|$, we obtain from (41) that $c/g \leq 9(e-1)/2000$. Then, we obtain from (3) that

$$\lambda \geq 0.144\eta,\ \chi = \frac{4}{K}, \tau \leq \frac{5\eta}{K},$$

which, in conjunction with (5), implies that

$$\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \leq \frac{140L(f_\beta(\mathbf{x}^0) - f_\beta(\mathbf{x}^*))}{(K+1)} + \frac{35}{K} \leq \mathcal{O}\left(\frac{1}{K}\right). \tag{42}$$

We choose $K = C\epsilon^{-1}$, where C is a positive constant. Then, based on the above inequality, we have, for C large enough, our Algorithm 1 achieves $\mathbb{E}\|f(\mathbf{x}^\zeta)\|^2 \leq \epsilon$, and the total number of function queries is

$$\left\lceil \frac{K}{q} \right\rceil nd + K|\mathcal{S}_2|d \leq \mathcal{O}\left(nd + \frac{nd}{\epsilon n^{1/3}} + \frac{n^{2/3}d}{\epsilon}\right) = \mathcal{O}\left(nd + \frac{n^{2/3}d}{\epsilon}\right) \leq \mathcal{O}\left(\frac{n^{2/3}d}{\epsilon}\right) \leq \mathcal{O}\left(\frac{d}{\epsilon^{5/3}}\right) \tag{43}$$

where the last two inequalities follow from the assumption that $n \leq K = C\epsilon^{-1}$.

Next, we suppose $n > K$. In this case, we have $|\mathcal{S}_1| = K$. Similarly to the case when $n \leq K$, we obtain

$$\begin{aligned} &c/g \leq \frac{9(e-1)}{2000},\ \lambda \geq 0.144\eta,\ \chi = \frac{9\sigma^2+4}{K} + \frac{18}{K^2}, \\ &\tau \leq \eta\left(\frac{18}{K^2} + \frac{9\sigma^2+4}{K}\right) + \frac{\eta}{K} \leq \frac{18\eta}{K^2} + \frac{(9\sigma^2+5)\eta}{K} \end{aligned} \tag{44}$$

which, in conjunction with (5), implies that

$$\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \leq \frac{140L(f_\beta(\mathbf{x}^0) - f_\beta(\mathbf{x}_\beta^*))}{(K+1)} + \frac{125}{K^2} + \frac{63\sigma^2+35}{K}$$

We choose $K = C\epsilon^{-1}$, where $C > 0$ is a positive constant. Then, based on the above inequality, we have, for C large enough, our Algorithm 1 achieves $\mathbb{E}\|f(\mathbf{x}^\zeta)\|^2 \leq \epsilon$, and the total number of function queries is

$$\begin{aligned} \left\lceil \frac{K}{q} \right\rceil Kd + K|\mathcal{S}_2|d &\leq Kd + K^{5/3}d + K^{5/3}d + Kd = 2K^{5/3}d + 2Kd \leq \mathcal{O}(K^{5/3}d) \\ &\leq \mathcal{O}(d\epsilon^{-5/3}) \leq \mathcal{O}(\epsilon^{-1}n^{2/3}d), \end{aligned}$$

where the last inequality follows from the assumption that $n > K \geq C\epsilon^{-1}$.

Combining the above two cases finishes the proof.

D.5. Proof of Corollary 2

We prove two cases with $n \leq \lceil (K/d)^{3/5} \rceil$ and $n > \lceil (K/d)^{3/5} \rceil$, separately.

First we suppose $n \leq \lceil (K/d)^{3/5} \rceil$, and thus we have $|\mathcal{S}_1| = n$ and $q = nd$. Based on (4), we have

$$c = 9dL^3\eta^2 \frac{(1+\theta)^q - 1}{\theta} \tag{45}$$

where $\theta = \eta g + 12\eta^2 dL^2$. Based on the parameter selection in (7), we have,

$$\theta = \frac{1}{2q} + \frac{3}{100n^{1/3}}\frac{1}{q} < \frac{1}{q} \text{ and } \theta > \frac{1}{2q}. \tag{46}$$

Combining (45) and (46) yields

$$c \leq 18(e-1)d\eta^2 qL^3 \leq \frac{9(e-1)L}{200n^{1/3}}. \tag{47}$$

Since $g = 4000d\eta^2 qL^3$, we obtain from (47) that $c/g \leq 9(e-1)/2000$, which, in conjunction with (3) and (7), implies that

$$\begin{aligned} &\lambda \geq 0.144\eta,\ \chi = \frac{4n^{2/3}}{K} \\ &\tau \leq \frac{4n^{2/3}\eta}{K} + \frac{n^{2/3}\eta}{K} < \frac{5\eta n^{2/3}}{K}, \end{aligned}$$

which, in conjunction with (5), implies that

$$\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \leq \frac{140Ldn^{2/3}(f_\beta(\mathbf{x}^0) - f_\beta(\mathbf{x}_\beta^*))}{(K+1)} + \frac{35dn^{2/3}}{K} \leq \mathcal{O}\left(\frac{dn^{2/3}}{K}\right). \tag{48}$$

Let $K = \lceil Cdn^{2/3}\epsilon^{-1} \rceil$ for a constant $C > 0$, which, in conjunction with the assumption that $n \leq \lceil (K/d)^{3/5} \rceil$, implies that $n \leq \Theta(\epsilon^{-1})$. Then, we have, for C large enough, $\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \leq \epsilon$, and the number of function queries is

$$\left\lceil \frac{K}{q} \right\rceil nd + K|\mathcal{S}_2|d \leq \mathcal{O}\left(nd + \frac{nd}{\epsilon n^{1/3}} + \frac{n^{2/3}d}{\epsilon}\right) = \mathcal{O}\left(nd + \frac{n^{2/3}d}{\epsilon}\right) \leq \mathcal{O}\left(\frac{n^{2/3}d}{\epsilon}\right) \leq \mathcal{O}\left(\frac{d}{\epsilon^{5/3}}\right) \tag{49}$$

where the last two inequalities follow from the assumption that $n \leq \lceil (K/d)^{3/5} \rceil \leq \mathcal{O}(\epsilon^{-1})$.

Next, we suppose $n > \lceil (K/d)^{3/5} \rceil$, and thus we have $|\mathcal{S}_1| = \lceil (K/d)^{3/5} \rceil$. Similarly to the case when $n \leq \lceil (K/d)^{3/5} \rceil$, we obtain

$$\begin{aligned} &c/g \leq \frac{9(e-1)}{2000},\ \lambda \geq 0.144\eta,\ \chi \leq \frac{(9\sigma^2+4)d^{3/5}}{K^{3/5}} + \frac{18d^{6/5}}{K^{6/5}}, \\ &\tau \leq \eta\left(\frac{(9\sigma^2+5)d^{3/5}}{K^{3/5}} + \frac{18d^{6/5}}{K^{6/5}}\right) \end{aligned} \tag{50}$$

which, in conjunction with (5), implies that

$$\begin{aligned} \mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \leq& \frac{140Ld|\mathcal{S}_1|^{2/3}(f_\beta(\mathbf{x}^0) - f_\beta(\mathbf{x}^*_\beta))}{K} + \frac{(63\sigma^2+35)d|\mathcal{S}_1|^{2/3}}{K} + \frac{125d}{K|\mathcal{S}_1|^{1/3}} \\ \leq& \mathcal{O}\left(\frac{d|\mathcal{S}_1|^{2/3}}{K}\right) \end{aligned} \tag{51}$$

where the first inequality follows from $|\mathcal{S}_1| = \lceil (K/d)^{3/5} \rceil$. Let $K = Cd\epsilon^{-5/3}$, where $C > 0$ is a large constant. Then, using (51) , we have, for C large enough, $\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \leq \epsilon$, and thus the number of function queries is

$$\left\lceil \frac{K}{q} \right\rceil |\mathcal{S}_1|d + K \leq \mathcal{O}\left(|\mathcal{S}_1|d + \frac{K}{q}|\mathcal{S}_1|d + K\right) \leq \mathcal{O}(K^{3/5}d^{2/5} + K) \leq \mathcal{O}(d\epsilon^{-5/3}) \leq \mathcal{O}\left(\frac{n^{2/3}d}{\epsilon}\right) \tag{52}$$

where the last two inequalities follow from $K = Cd\epsilon^{-5/3}$ and the assumption that $n > \lceil (K/d)^{3/5} \rceil = C^{3/5}\epsilon^{-1}$.

Combining the above two cases finishes the proof.

E. Proof for ZO-SVRG-Coord

E.1. Proof of Lemma 2

For any $qk_0 \leq k \leq \min\{q(k_0+1)-1, qh\},\ k_0 = 0, ..., h$, based on (11), we obtain

$$\begin{aligned} &\mathbb{E}\|\mathbf{v}^k - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)\|^2 \\ &\quad \leq 2\mathbb{E}\|\mathbf{v}^k - \mathbf{v}^{qk_0} - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0}))\|^2 + 2\mathbb{E}\|\mathbf{v}^{qk_0} - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0})\|^2. \end{aligned} \tag{53}$$

To simplify notation, we denote

$$G_{\mathcal{S}_2}(\mathbf{x}) = \frac{1}{|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|} \hat{\nabla}_{\text{coord}} f_{a_j}(\mathbf{x}) \text{ and } H_j(\mathbf{x}) = \hat{\nabla}_{\text{coord}} f_{a_j}(\mathbf{x}^t), \tag{54}$$

which, in conjunction with (53), implies that

$$\begin{aligned} \mathbb{E}\|\mathbf{v}^k - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)\|^2 =& \underbrace{2\,\mathbb{E}\left(\mathbb{E}\|G_{\mathcal{S}_2}(\mathbf{x}^k) - G_{\mathcal{S}_2}(\mathbf{x}^{qk_0}) - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0}))\|^2 \,|\, \mathbf{x}^0, ..., \mathbf{x}^k\right)}_{P} \\ &+ 2\mathbb{E}\|\mathbf{v}^{qk_0} - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0})\|^2. \end{aligned} \tag{55}$$

Conditioned on $\mathbf{x}^0, ..., \mathbf{x}^k$, we next provide an upper bound on the conditional expectation term P in (55). Using the shorthand $\mathbb{E}_k(\cdot)$ to denote $\mathbb{E}(\cdot \,|\, \mathbf{x}^1, ..., \mathbf{x}^k)$, we have

$$\begin{aligned}
&\mathbb{E}_k\|G_{\mathcal{S}_2}(\mathbf{x}^k) - G_{\mathcal{S}_2}(\mathbf{x}^{qk_0}) - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0}))\|^2 \\
=&\mathbb{E}_k\left\| \frac{1}{|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|} \left(H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0}) - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0}))\right)\right\|^2 \\
=&\frac{1}{|\mathcal{S}_2|^2}\sum_{j=1}^{|\mathcal{S}_2|} \mathbb{E}_k\left\| H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0}) - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0}))\right\|^2 \\
&- 2\sum_{i\neq j} \mathbb{E}_k \big\langle H_i(\mathbf{x}^k) - H_i(\mathbf{x}^{qk_0}) - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0})), \\
&\qquad\qquad H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0}) - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0}))\big\rangle \\
\overset{\text{(i)}}{=}&\frac{1}{|\mathcal{S}_2|^2}\sum_{j=1}^{|\mathcal{S}_2|} \mathbb{E}_k\left\| H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0}) - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0}))\right\|^2
\end{aligned} \tag{56}$$

where (i) follows from the facts that a_i is independent of a_j for any $i \neq j$, $\mathbb{E}_k(H_j(\mathbf{x}^k)) = \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)$ and $\mathbb{E}_k(H_j(\mathbf{x}^{qk_0})) = \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0})$. Then, we further simplify (56) to

$$\begin{aligned}
&\frac{1}{|\mathcal{S}_2|}\mathbb{E}_k\left\| H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0}) - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0}))\right\|^2 \\
&= \frac{1}{|\mathcal{S}_2|}\mathbb{E}_k\left\| H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0})\right\|^2 + \frac{1}{|\mathcal{S}_2|}\|\hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0})\|^2 \\
&\quad - \frac{2}{|\mathcal{S}_2|}\mathbb{E}_k\left\langle H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0}), \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0})\right\rangle \\
&\overset{\text{(i)}}{=} \frac{1}{|\mathcal{S}_2|}\mathbb{E}_k\left\| H_j(\mathbf{x}^k) - H_j(\mathbf{x}^{qk_0})\right\|^2 - \frac{1}{|\mathcal{S}_2|}\|\hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0})\|^2 \\
&\overset{\text{(ii)}}{\leq} \frac{3}{|\mathcal{S}_2|}\mathbb{E}_k\left\| \hat{\nabla}_{\text{coord}} f_{a_j}(\mathbf{x}^k) - \nabla f_{a_j}(\mathbf{x}^k)\right\|^2 + \frac{3}{|\mathcal{S}_2|}\mathbb{E}_k\left\| \nabla f_{a_j}(\mathbf{x}^k) - \nabla f_{a_j}(\mathbf{x}^{qk_0})\right\|^2 \\
&\quad + \frac{3}{|\mathcal{S}_2|}\mathbb{E}_k\left\| \hat{\nabla}_{\text{coord}} f_{a_j}(\mathbf{x}^{qk_0}) - \nabla f_{a_j}(\mathbf{x}^{qk_0})\right\|^2 \\
&\overset{\text{(iii)}}{\leq} \frac{6L^2 d\delta^2}{|\mathcal{S}_2|} + \frac{3L^2}{|\mathcal{S}_2|}\|\mathbf{x}^k - \mathbf{x}^{qk_0}\|^2
\end{aligned} \tag{57}$$

where (i) follows from the fact that $\mathbb{E}_k(H_j(\mathbf{x}^k)) = \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)$ and $\mathbb{E}_k(H_j(\mathbf{x}^{qk_0})) = \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1})$, (ii) follows from the inequality that $\|\mathbf{a}+\mathbf{b}+\mathbf{c}\|^2 \leq 3(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + \|\mathbf{c}\|^2)$, and (iii) follows from Lemma 3 and Assumption 1. Combining (55), (57), Lemma 4 and unconditioned on $\mathbf{x}^0, ..., \mathbf{x}^k$, we have

$$\mathbb{E}\|\mathbf{v}^k - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)\|^2 \leq \frac{12L^2 d\delta^2}{|\mathcal{S}_2|} + \frac{6L^2}{|\mathcal{S}_2|}\|\mathbf{x}^k - \mathbf{x}^{qk_0}\|^2 + \frac{6I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right),$$

which finishes the proof.

E.2. Proof of Theorem 2

Since $K = qh$ and $\nabla f_\beta(\mathbf{x})$ is L-Lipschitz, we have, for $qm \leq k \leq q(m+1) - 1, m = 0, ..., h-1$

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k\rangle + \frac{L\eta^2}{2}\|\mathbf{v}^k\|^2 = f(\mathbf{x}^k) - \eta\langle \nabla f(\mathbf{x}^k), \mathbf{v}^k\rangle + \frac{L\eta^2}{2}\|\mathbf{v}^k\|^2.$$

Taking the expectation over the above inequality and noting that $\mathbb{E}(\mathbf{v}^k \mid \mathbf{x}^0, ..., \mathbf{x}^k) = \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qm}) + \mathbf{v}^{qm}$, we have

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}^{k+1}) \leq & \mathbb{E}f(\mathbf{x}^k) - \eta\mathbb{E}\langle \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qm}) + \mathbf{v}^{qm}, \nabla f(\mathbf{x}^k)\rangle + \frac{L\eta^2}{2}\mathbb{E}\|\mathbf{v}^k\|^2 \\
\leq & \mathbb{E}f(\mathbf{x}^k) - \eta\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \eta\mathbb{E}\langle \nabla f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) + \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qm}) - \mathbf{v}^{qm}, \nabla f(\mathbf{x}^k)\rangle + \frac{L\eta^2}{2}\mathbb{E}\|\mathbf{v}^k\|^2 \\
\leq & \mathbb{E}f(\mathbf{x}^k) - \eta\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \frac{\eta}{2}\mathbb{E}\|\nabla f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) + \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qm}) - \mathbf{v}^{qm}\|^2 \\
& + \frac{\eta}{2}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \frac{L\eta^2}{2}\mathbb{E}\|\mathbf{v}^k\|^2 \\
\leq & \mathbb{E}f(\mathbf{x}^k) - \frac{\eta}{2}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \eta\mathbb{E}\|\nabla f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)\|^2 + \eta\mathbb{E}\|\hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qm}) - \mathbf{v}^{qm}\|^2 + \frac{L\eta^2}{2}\mathbb{E}\|\mathbf{v}^k\|^2 \\
\overset{\text{(i)}}{\leq} & \mathbb{E}f(\mathbf{x}^k) - \frac{\eta}{2}\|\nabla f(\mathbf{x}^k)\|^2 + \eta L^2 d\delta^2 + \frac{3\eta I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right) + \frac{L\eta^2}{2}\mathbb{E}\|\mathbf{v}^k\|^2
\end{aligned} \tag{58}$$

where (i) follows from Lemmas 3 and 4.

We introduce a Lyapunov function $R_k^m = \mathbb{E}\left(f(\mathbf{x}^k) + c_k^m\|\mathbf{x}^k - \mathbf{x}^{qm}\|^2\right)$ for $qm \leq k \leq q(m+1), m = 0, ..., h-1$, where $\{c_k^m\}$ are constants such that $c_{q(m+1)}^m = 0$. Then, we obtain that for any $qm \leq k \leq q(m+1) - 1$

$$\begin{aligned}
R_{k+1}^m = & \mathbb{E}\left(f(\mathbf{x}^{k+1}) + c_{k+1}^m\|\mathbf{x}^{k+1} - \mathbf{x}^k + \mathbf{x}^k - \mathbf{x}^{qm}\|^2\right) \\
\leq & \mathbb{E}f(\mathbf{x}^{k+1}) + c_{k+1}^m\eta^2\mathbb{E}\|\mathbf{v}^k\|^2 + c_{k+1}^m\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{qm}\|^2 - 2c_{k+1}^m\eta\mathbb{E}\langle \mathbf{v}^k, \mathbf{x}^k - \mathbf{x}^{qm}\rangle \\
\overset{\text{(i)}}{=} & \mathbb{E}f(\mathbf{x}^{k+1}) + c_{k+1}^m\eta^2\mathbb{E}\|\mathbf{v}^k\|^2 + c_{k+1}^m\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{qm}\|^2 - 2c_{k+1}^m\eta\mathbb{E}\langle \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qm}) + \mathbf{v}^{qm}, \mathbf{x}^k - \mathbf{x}^{qm}\rangle \\
\leq & \mathbb{E}f(\mathbf{x}^{k+1}) + c_{k+1}^m\eta^2\mathbb{E}\|\mathbf{v}^k\|^2 + (c_{k+1}^m + c_{k+1}^m\eta g)\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{qm}\|^2 \\
& + \frac{2c_{k+1}^m\eta}{g}\mathbb{E}\left(\|\hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)\|^2 + \|\hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qm}) - \mathbf{v}^{qm}\|^2\right) \\
\overset{\text{(ii)}}{\leq} & \mathbb{E}f(\mathbf{x}^{k+1}) + c_{k+1}^m\eta^2\mathbb{E}\|\mathbf{v}^k\|^2 + (c_{k+1}^m + c_{k+1}^m\eta g)\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{qm}\|^2 \\
& + \frac{4c_{k+1}^m\eta}{g}\left(\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + L^2 d\delta^2\right) + \frac{12c_{k+1}^m\eta}{g}\frac{I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right).
\end{aligned} \tag{59}$$

where (i) follows from the definition of $\mathbf{v}^k$ and (ii) follows from Lemma 2. Combining (58) and (59), we obtain

$$\begin{aligned}
R_{k+1}^m \leq & \mathbb{E}f(\mathbf{x}^k) - \left(\frac{\eta}{2} - \frac{4c_{k+1}^m\eta}{g}\right)\|\nabla f(\mathbf{x}^k)\|^2 + \left(\frac{L}{2} + c_{k+1}^m\right)\eta^2\mathbb{E}\|\mathbf{v}^k\|^2 + (c_{k+1}^m + c_{k+1}^m\eta g)\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{qm}\|^2 \\
& + \frac{4c_{k+1}^m\eta}{g}L^2 d\delta^2 + \eta L^2 d\delta^2 + \left(\frac{12c_{k+1}^m\eta}{g} + 3\eta\right)\frac{I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right).
\end{aligned} \tag{60}$$

Based on Lemma 2, we obtain

$$\begin{aligned}
\mathbb{E}\|\mathbf{v}^k\|^2 \leq & 3\mathbb{E}\|\mathbf{v}^k - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)\|^2 + 3\mathbb{E}\|\nabla f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)\|^2 + 3\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\
\leq & \frac{36L^2 d\delta^2}{|\mathcal{S}_2|} + \frac{18L^2}{|\mathcal{S}_2|}\|\mathbf{x}^k - \mathbf{x}^{qm}\|^2 + \frac{18I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right) + 3L^2 d\delta^2 + 3\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2,
\end{aligned}$$

which, in conjunction with (60), implies that

$$\begin{aligned}
R_{k+1}^m \leq & \mathbb{E}f(\mathbf{x}^k) - \left(\frac{\eta}{2} - \frac{4c_{k+1}^m\eta}{g} - 3\left(\frac{L}{2} + c_{k+1}^m\right)\eta^2\right)\|\nabla f(\mathbf{x}^k)\|^2 \\
& + \left(c_{k+1}^m + c_{k+1}^m\eta g + \frac{18L^2}{|\mathcal{S}_2|}\left(\frac{L}{2} + c_{k+1}^m\right)\eta^2\right)\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{qm}\|^2 \\
& + \left(\frac{12c_{k+1}^m\eta}{g} + 3\eta + 18\left(\frac{L}{2} + c_{k+1}^m\right)\eta^2\right)\frac{I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right) \\
& + \frac{4c_{k+1}^m\eta}{g}L^2 d\delta^2 + \eta L^2 d\delta^2 + \left(\frac{36}{|\mathcal{S}_2|} + 3\right)\left(\frac{L}{2} + c_{k+1}^m\right)\eta^2 L^2 d\delta^2.
\end{aligned} \tag{61}$$

Let $c_k^m := (1+\theta)c_{k+1}^m + \frac{9L^3\eta^2}{|\mathcal{S}_2|}$, where $\theta = \eta g + \frac{18L^2\eta^2}{|\mathcal{S}_2|}$. Then, we rewrite (61) as

$$\begin{aligned} R_{k+1}^m \leq & R_k^m - \left(\frac{\eta}{2} - \frac{4c_{k+1}^m\eta}{g} - 3\left(\frac{L}{2} + c_{k+1}^m\right)\eta^2\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\ & + \left(\frac{12c_{k+1}^m\eta}{g} + 3\eta + 18\left(\frac{L}{2} + c_{k+1}^m\right)\eta^2\right)\frac{I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2d\delta^2 + \sigma^2\right) \\ & + \frac{4c_{k+1}^m\eta}{g}L^2d\delta^2 + \eta L^2 d\delta^2 + \left(\frac{36}{|\mathcal{S}_2|} + 3\right)\left(\frac{L}{2} + c_{k+1}^m\right)\eta^2L^2d\delta^2. \end{aligned} \tag{62}$$

Note that for $qm \leq k \leq q(m+1) - 1, m = 0, ..., h-1$

$$c_k^m \leq c = \frac{9L^3\eta^2}{|\mathcal{S}_2|}\frac{(1+\theta)^q - 1}{\theta}.$$

To simplify notation, we define

$$\begin{aligned} \lambda &= \frac{\eta}{2} - \frac{4c\eta}{g} - 3\left(\frac{L}{2} + c\right)\eta^2 \\ \chi &= \left(\frac{12c\eta}{g} + 3\eta + 18\left(\frac{L}{2} + c\right)\eta^2\right)\frac{I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2d\delta^2 + \sigma^2\right) \\ \tau &= \left(\frac{4c}{g} + 1\right)\eta L^2 d\delta^2 + \left(\frac{36}{|\mathcal{S}_2|} + 3\right)\left(\frac{L}{2} + c\right)\eta^2L^2d\delta^2, \end{aligned} \tag{63}$$

which, in conjunction with (62), implies that

$$R_{k+1}^m \leq R_k^m - \lambda\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \tau + \chi.$$

Telescoping the above inequality over k from qm to $q(m+1) - 1$ and noting that $R_{qm}^m = \mathbb{E}f(\mathbf{x}^{qm})$ and $R_{q(m+1)}^m = \mathbb{E}f(\mathbf{x}^{q(m+1)})$, we obtain

$$\mathbb{E}f(\mathbf{x}^{q(m+1)}) \leq \mathbb{E}f(\mathbf{x}^{qm}) - \lambda\sum_{k=qm}^{q(m+1)-1}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + q\tau + q\chi$$

Then, telescoping the above inequality over m from 0 to $h-1$, we obtain

$$\mathbb{E}f(\mathbf{x}^K) \leq \mathbb{E}f(\mathbf{x}^0) - \lambda\sum_{k=0}^{K}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + K\tau + K\chi,$$

which, in conjunction with the definition of $\mathbf{x}^\zeta$, implies that

$$\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \leq \frac{\Delta}{\lambda K} + \frac{\tau + \chi}{\lambda}. \tag{64}$$

where $\Delta := f(\mathbf{x}^0) - f(\mathbf{x}^*)$.

Let $g = 1/(2\eta q)$. Then, based on the selected parameters in (9) and the definition of θ, we have

$$\frac{1}{2q} < \theta \leq \frac{1}{2q} + \frac{2}{25q}\frac{1}{q} \leq \frac{1}{q},$$

which, in conjunction with the definition of c, implies that

$$c \leq \frac{18(e-1)L^3\eta^2q}{|\mathcal{S}_2|} \leq \frac{2(e-1)L}{25q}, \tag{65}$$

and $c/g \leq 0.02$.

Next, we prove two cases when $n \leq K$ and $n > K$, separately. First suppose $n \leq K$. In such a case, we have $|\mathcal{S}_1| = n$ and $q = \lceil n^{1/3} \rceil$. Then, based on (63), (9) and (65), we obtain

$$\lambda \geq 0.22\eta,\ \chi = 0,\ \tau < \frac{5\eta}{K},$$

which, in conjunction with (64), yields

$$\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \leq \frac{69\Delta + 23}{K} \leq \mathcal{O}\left(\frac{1}{K}\right). \tag{66}$$

Let $K = C\epsilon^{-1}$, where C is a constant. Then, we have, for C large enough, $\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \leq \epsilon$, and the number of function queries is

$$\left\lceil \frac{K}{q} \right\rceil nd + K|\mathcal{S}_2|d \leq \mathcal{O}\left(nd + \frac{dn^{2/3}}{\epsilon}\right) \leq \mathcal{O}\left(\frac{dn^{2/3}}{\epsilon}\right) \leq \mathcal{O}\left(\min\left\{\frac{n^{2/3}d}{\epsilon}, \frac{d}{\epsilon^{5/3}}\right\}\right) \tag{67}$$

where the last two inequalities follow from the assumption that $n \leq K = C\epsilon^{-1}$.

Next, we suppose $n > K$. In this case, we obtain

$$\lambda \geq 0.22\eta,\ \chi \leq \frac{5\eta}{K}\left(\frac{2}{K} + \sigma^2\right),\ \tau \leq \frac{5\eta}{K}$$

which, in conjunction with (64), yields

$$\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \leq \frac{69\Delta + 23 + 23\sigma^2}{K} + \frac{46}{K^2} \tag{68}$$

Let $K = C\epsilon^{-1}$, where C is a constant. Then, we have, for C large enough, $\mathbb{E}\|\nabla f(\mathbf{x}^\zeta)\|^2 \leq \epsilon$, and the number of function queries is

$$\left\lceil \frac{K}{q} \right\rceil Kd + K|\mathcal{S}_2|d \leq \mathcal{O}\left(dK^{5/3}\right) \leq \mathcal{O}\left(\min\left\{\frac{n^{2/3}d}{\epsilon}, \frac{d}{\epsilon^{5/3}}\right\}\right) \tag{69}$$

where the last inequality follows from the assumption that $n > K = C\epsilon^{-1}$.

Combining the above two cases finish the proof.

F. Proofs for ZO-SPIDER-Coord

F.1. Auxiliary Lemma

The following lemma provides an upper bound on the error of $\mathbf{v}^k$ for estimating the second moment of $\|\mathbf{v}^k - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)\|$.

Lemma 6. *For any given $k_0 \leq \lfloor K/q \rfloor$ and $qk_0 \leq k \leq \min\{q(k_0+1)-1, K\}$, we have*

$$\mathbb{E}\|\mathbf{v}^k - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)\|^2 \leq \frac{3\eta^2 L^2}{|\mathcal{S}_2|} \sum_{t=qk_0}^{k-1} \mathbb{E}\|\mathbf{v}^t\|^2 + (k - qk_0)\frac{6L^2 d\delta^2}{|\mathcal{S}_2|} + \frac{3I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right). \tag{70}$$

where we define $\sum_{t=qk_0}^{qk_0-1} \mathbb{E}\|\mathbf{v}^t\|^2 = 0$ for simplicity.

Proof. First we consider the case when $k \geq qk_0 + 1$. For $qk_0 + 1 \leq m \leq k$, we have

$$\mathbf{v}^m - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) = \mathbf{v}^{qk_0} - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0}) + \sum_{t=qk_0+1}^{m} (\mathbf{v}^t - \mathbf{v}^{t-1} - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^t) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{t-1}))). \tag{71}$$

Recall that $\mathbf{v}^t$ is given by

$$\mathbf{v}^t = \frac{1}{|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|} \hat{\nabla}_{\text{coord}} f_{a_j}(\mathbf{x}^t) - \frac{1}{|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|} \hat{\nabla}_{\text{coord}} f_{a_j}(\mathbf{x}^{t-1}) + \mathbf{v}^{t-1}. \tag{72}$$

We then have for any $qk_0 \leq t \leq m$, $\mathbb{E}(\mathbf{v}^t - \mathbf{v}^{t-1} - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^t) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{t-1}))\,|\mathbf{x}^0, ..., \mathbf{x}^t) = 0$, which, in conjunction with (71), implies that the sequence $(\mathbf{v}^t - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^t), t = qk_0, ..., m)$ is a martingale. Then, based on the property of square-integrable martingales (Fang et al., 2018), we can obtain, for $qk_0 + 1 \leq m \leq k$,

$$\begin{aligned}\mathbb{E}\|\mathbf{v}^m - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^m)\|^2 =& \mathbb{E}\|\mathbf{v}^{qk_0} - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0})\|^2 \\ &+ \sum_{t=qk_0+1}^{m} \mathbb{E}\|\mathbf{v}^t - \mathbf{v}^{t-1} - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^t) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{t-1}))\|^2.\end{aligned}$$

The above equality further implies that

$$\begin{aligned}&\mathbb{E}\|\mathbf{v}^m - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^m)\|^2 \\ &= \mathbb{E}\|\mathbf{v}^m - \mathbf{v}^{m-1} - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1}))\|^2 + \mathbb{E}\|\mathbf{v}^{m-1} - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1})\|^2. \end{aligned} \tag{73}$$

Based on (72) and using the same notations as in (54), we have $\mathbf{v}^m - \mathbf{v}^{m-1} = G_{\mathcal{S}_2}(\mathbf{x}^m) - G_{\mathcal{S}_2}(\mathbf{x}^{m-1})$, which, in conjunction with (73), implies

$$\begin{aligned}\mathbb{E}\|\mathbf{v}^m - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^m)\|^2 =& \underbrace{\mathbb{E}\left(\mathbb{E}\|G_{\mathcal{S}_2}(\mathbf{x}^m) - G_{\mathcal{S}_2}(\mathbf{x}^{m-1}) - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1}))\|^2 \,|\, \mathbf{x}^0, ..., \mathbf{x}^m\right)}_{Q} \\ &+ \mathbb{E}\|\mathbf{v}^{m-1} - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1})\|^2. \end{aligned} \tag{74}$$

Conditioned on $\mathbf{x}^0, ..., \mathbf{x}^m$, we next provide an upper bound on the conditional expectation term Q in (74). Using the shorthand $\mathbb{E}_m(\cdot)$ to denote $\mathbb{E}(\cdot \,|\, \mathbf{x}^1, ..., \mathbf{x}^m)$, we have

$$\begin{aligned}&\mathbb{E}_m\|G_{\mathcal{S}_2}(\mathbf{x}^m) - G_{\mathcal{S}_2}(\mathbf{x}^{m-1}) - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1}))\|^2 \\ &\quad= \mathbb{E}_m\left\|\frac{1}{|\mathcal{S}_2|}\sum_{j=1}^{|\mathcal{S}_2|}\left(H_j(\mathbf{x}^m) - H_j(\mathbf{x}^{m-1}) - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1}))\right)\right\|^2 \\ &\quad= \frac{1}{|\mathcal{S}_2|^2}\sum_{j=1}^{|\mathcal{S}_2|}\mathbb{E}_m\left\|H_j(\mathbf{x}^m) - H_j(\mathbf{x}^{m-1}) - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1}))\right\|^2 \\ &\qquad- 2\sum_{i\neq j}\mathbb{E}_m\big\langle H_i(\mathbf{x}^m) - H_i(\mathbf{x}^{m-1}) - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1})), \\ &\qquad\qquad H_j(\mathbf{x}^m) - H_j(\mathbf{x}^{m-1}) - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1}))\big\rangle \\ &\quad\overset{\text{(i)}}{=} \frac{1}{|\mathcal{S}_2|^2}\sum_{j=1}^{|\mathcal{S}_2|}\mathbb{E}_m\left\|H_j(\mathbf{x}^m) - H_j(\mathbf{x}^{m-1}) - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1}))\right\|^2\end{aligned}$$

where (i) follows from the facts that a_i is independent of a_j for any $i \neq j$, $\mathbb{E}_m(H_j(\mathbf{x}^m)) = \hat{\nabla}_{\text{coord}} f(\mathbf{x}^m)$, and $\mathbb{E}_m(H_j(\mathbf{x}^{m-1})) = \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1})$. Then, we further simplify the above equation to

$$\begin{aligned}&\frac{1}{|\mathcal{S}_2|}\mathbb{E}_m\left\|H_j(\mathbf{x}^m) - H_j(\mathbf{x}^{m-1}) - (\hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1}))\right\|^2 \\ &\quad= \frac{1}{|\mathcal{S}_2|}\mathbb{E}_m\left\|H_j(\mathbf{x}^m) - H_j(\mathbf{x}^{m-1})\right\|^2 + \frac{1}{|\mathcal{S}_2|}\|\hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1})\|^2 \\ &\qquad- \frac{2}{|\mathcal{S}_2|}\mathbb{E}_m\left\langle H_j(\mathbf{x}^m) - H_j(\mathbf{x}^{m-1}), \hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1})\right\rangle \\ &\quad\overset{\text{(i)}}{=} \frac{1}{|\mathcal{S}_2|}\mathbb{E}_m\left\|H_j(\mathbf{x}^m) - H_j(\mathbf{x}^{m-1})\right\|^2 - \frac{1}{|\mathcal{S}_2|}\|\hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1})\|^2 \\ &\quad\leq \frac{1}{|\mathcal{S}_2|}\mathbb{E}_m\left\|H_j(\mathbf{x}^m) - H_j(\mathbf{x}^{m-1})\right\|^2\end{aligned}$$

$$\overset{\text{(ii)}}{\leq} \frac{3}{|\mathcal{S}_2|}\mathbb{E}_m\left\|\hat{\nabla}_{\text{coord}} f_{a_j}(\mathbf{x}^m) - \nabla f_{a_j}(\mathbf{x}^m)\right\|^2 + \frac{3}{|\mathcal{S}_2|}\mathbb{E}_m\left\|\nabla f_{a_j}(\mathbf{x}^m) - \nabla f_{a_j}(\mathbf{x}^{m-1})\right\|^2$$
$$+ \frac{3}{|\mathcal{S}_2|}\mathbb{E}_m\left\|\hat{\nabla}_{\text{coord}} f_{a_j}(\mathbf{x}^{m-1}) - \nabla f_{a_j}(\mathbf{x}^{m-1})\right\|^2$$
$$\overset{\text{(iii)}}{\leq} \frac{6L^2 d\delta^2}{|\mathcal{S}_2|} + \frac{3L^2}{|\mathcal{S}_2|}\|\mathbf{x}^m - \mathbf{x}^{m-1}\|^2 = \frac{6L^2 d\delta^2}{|\mathcal{S}_2|} + \frac{3\eta^2 L^2}{|\mathcal{S}_2|}\|\mathbf{v}^{m-1}\|^2 \tag{75}$$

where (i) follows from the fact that $\mathbb{E}_m(H_j(\mathbf{x}^m)) = \hat{\nabla}_{\text{coord}} f(\mathbf{x}^m)$ and $\mathbb{E}_m(H_j(\mathbf{x}^{m-1})) = \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1})$, (ii) follows from the inequality that $\|\mathbf{a}+\mathbf{b}+\mathbf{c}\|^2 \leq 3(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + \|\mathbf{c}\|^2)$, and (iii) follows from Lemma 3 and Assumption 1. Combining (73) and (75) and unconditioned on $\mathbf{x}^0, ..., \mathbf{x}^m$, we obtain

$$\mathbb{E}\|\mathbf{v}^m - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^m)\|^2 \leq \frac{6L^2 d\delta^2}{|\mathcal{S}_2|} + \frac{3\eta^2 L^2}{|\mathcal{S}_2|}\|\mathbf{v}^{m-1}\|^2 + \mathbb{E}\|\mathbf{v}^{m-1} - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{m-1})\|^2. \tag{76}$$

Telescoping the above inequality over m from qk_0+1 to k, we obtain

$$\mathbb{E}\|\mathbf{v}^k - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)\|^2 \leq \frac{3\eta^2 L^2}{|\mathcal{S}_2|}\sum_{t=qk_0+1}^{k}\mathbb{E}\|\mathbf{v}^{t-1}\|^2 + (k-qk_0)\frac{6L^2 d\delta^2}{|\mathcal{S}_2|}$$
$$+ \mathbb{E}\|\mathbf{v}^{qk_0} - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0})\|^2. \tag{77}$$

Using Lemma 4 and (77) yields (70). For the case when $k = qk_0$, it can be checked that (70) also holds. □

F.2. Proof of Theorem 3

Noting that $f(\cdot)$ has a L-Lipschitz gradient, we have, for any given $k_0 \leq \lfloor K/q \rfloor$ and $qk_0 \leq m \leq \min\{q(k_0+1)-1, K\}$,

$$\begin{aligned} f(\mathbf{x}^{m+1}) &\leq f(\mathbf{x}^m) + \langle \nabla f(\mathbf{x}^m), \mathbf{x}^{m+1} - \mathbf{x}^m\rangle + \frac{L}{2}\|\mathbf{x}^{m+1} - \mathbf{x}^m\|^2 \\ &= f(\mathbf{x}^m) - \eta\langle \nabla f(\mathbf{x}^m) - \mathbf{v}^m, \mathbf{v}^m\rangle - \eta\|\mathbf{v}^m\|^2 + \frac{L\eta^2}{2}\|\mathbf{v}^m\|^2 \\ &\leq f(\mathbf{x}^m) + \frac{\eta}{2}\|\nabla f(\mathbf{x}^m) - \mathbf{v}^m\|^2 + \frac{\eta}{2}\|\mathbf{v}^m\|^2 - \eta\|\mathbf{v}^m\|^2 + \frac{L\eta^2}{2}\|\mathbf{v}^m\|^2 \\ &= f(\mathbf{x}^m) + \frac{\eta}{2}\|\nabla f(\mathbf{x}^m) - \mathbf{v}^m\|^2 - \Big(\frac{\eta}{2} - \frac{L\eta^2}{2}\Big)\|\mathbf{v}^m\|^2. \end{aligned}$$

Taking the expectation over the above inequality yields

$$\begin{aligned} \mathbb{E}f(\mathbf{x}^{m+1}) &\leq \mathbb{E}f(\mathbf{x}^m) + \eta\left(\mathbb{E}\|\hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) - \mathbf{v}^m\|^2 + \mathbb{E}\|\nabla f(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^m)\|^2\right) - \Big(\frac{\eta}{2} - \frac{L\eta^2}{2}\Big)\mathbb{E}\|\mathbf{v}^m\|^2 \\ &\leq \mathbb{E}f(\mathbf{x}^m) + \eta\left(\mathbb{E}\|\hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) - \mathbf{v}^m\|^2 + L^2 d\delta^2\right) - \Big(\frac{\eta}{2} - \frac{L\eta^2}{2}\Big)\mathbb{E}\|\mathbf{v}^m\|^2, \end{aligned}$$

which, in conjunction with Lemma 6, implies that

$$\mathbb{E}f(\mathbf{x}^{m+1}) \leq \mathbb{E}f(\mathbf{x}^m) + \frac{3\eta^3 L^2}{|\mathcal{S}_2|}\sum_{t=qk_0}^{m-1}\mathbb{E}\|\mathbf{v}^t\|^2 + (m-qk_0)\frac{6\eta L^2 d\delta^2}{|\mathcal{S}_2|} + \frac{3\eta I(|\mathcal{S}_1|<n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right) + \eta L^2 d\delta^2$$
$$- \Big(\frac{\eta}{2} - \frac{L\eta^2}{2}\Big)\mathbb{E}\|\mathbf{v}^m\|^2. \tag{78}$$

To simplify notation, we define

$$\pi(\mathcal{S}_1, \delta) = \frac{3I(|\mathcal{S}_1|<n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right) + L^2 d\delta^2. \tag{79}$$

Then, telescoping (78) over m from qk_0 to k yields

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}^{k+1}) \leq & \mathbb{E}f(\mathbf{x}^{qk_0}) + \frac{3L^2\eta^3}{|\mathcal{S}_2|}\sum_{t_1=qk_0}^{k}\sum_{t_2=qk_0}^{t_1-1}\mathbb{E}\|\mathbf{v}^{t_2}\|^2 + \sum_{t_1=qk_0}^{k}(t_1-qk_0)\frac{6\eta L^2 d\delta^2}{|\mathcal{S}_2|} \\
& + (k-qk_0+1)\eta\pi(\mathcal{S}_1,\delta) - \Big(\frac{\eta}{2}-\frac{L\eta^2}{2}\Big)\sum_{t_1=qk_0}^{k}\mathbb{E}\|\mathbf{v}^{k}\|^2. \\
\overset{\text{(i)}}{\leq} & \mathbb{E}f(\mathbf{x}^{qk_0}) + \frac{3L^2\eta^3}{|\mathcal{S}_2|}\sum_{t_1=qk_0}^{k}\sum_{t_2=qk_0}^{k}\mathbb{E}\|\mathbf{v}^{t_2}\|^2 + \frac{(k-qk_0)(k-qk_0+1)}{|\mathcal{S}_2|}3\eta L^2 d\delta^2 \\
& + (k-qk_0+1)\eta\pi(\mathcal{S}_1,\delta) - \Big(\frac{\eta}{2}-\frac{L\eta^2}{2}\Big)\sum_{t_1=qk_0}^{k}\mathbb{E}\|\mathbf{v}^{k}\|^2. \\
\leq & \mathbb{E}f(\mathbf{x}^{qk_0}) + \frac{3L^2\eta^3(k-k_0+1)}{|\mathcal{S}_2|}\sum_{t=qk_0}^{k}\mathbb{E}\|\mathbf{v}^{t}\|^2 + \frac{(k-qk_0)(k-qk_0+1)}{|\mathcal{S}_2|}3\eta L^2 d\delta^2 \\
& + (k-qk_0+1)\eta\pi(\mathcal{S}_1,\delta) - \Big(\frac{\eta}{2}-\frac{L\eta^2}{2}\Big)\sum_{t_1=qk_0}^{k}\mathbb{E}\|\mathbf{v}^{k}\|^2. \\
\leq & \mathbb{E}f(\mathbf{x}^{qk_0}) - \left(\frac{\eta}{2}-\frac{\eta^2 L}{2}-\frac{3L^2\eta^3(k-k_0+1)}{|\mathcal{S}_2|}\right)\sum_{t=qk_0}^{k}\mathbb{E}\|\mathbf{v}^{t}\|^2 \\
& + \frac{(k-qk_0)(k-qk_0+1)}{|\mathcal{S}_2|}3\eta L^2 d\delta^2 + (k-qk_0+1)\eta\pi(\mathcal{S}_1,\delta).
\end{aligned} \tag{80}$$

where (i) follows from the fact that $\sum_{t_2=qk_0}^{t_1-1}\mathbb{E}\|\mathbf{v}^{t_2}\|^2 \leq \sum_{t_2=qk_0}^{k}\mathbb{E}\|\mathbf{v}^{t_2}\|^2$ for $t_1-1 \leq k-1 < k$. Without loss of generality we suppose $k^\star q < K \leq (k^\star+1)q-1$, where $k^\star = \lfloor K/q \rfloor$. Then, based on (80), we have, after K iterations,

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}^{K}) \leq & \mathbb{E}f(\mathbf{x}^{0}) - \left(\frac{\eta}{2}-\frac{\eta^2 L}{2}-\frac{3L^2\eta^3(k-k^\star)}{|\mathcal{S}_2|}\right)\sum_{t=qk^\star}^{K-1}\mathbb{E}\|\mathbf{v}^{t}\|^2 + \frac{(K-qk^\star)^2}{|\mathcal{S}_2|}3\eta L^2 d\delta^2 \\
& + (K-qk^\star)\eta\pi(\mathcal{S}_1,\delta) + \sum_{t=1}^{k^\star}\Big(\mathbb{E}f(\mathbf{x}^{tq}) - \mathbb{E}f(\mathbf{x}^{(t-1)q})\Big).
\end{aligned} \tag{81}$$

The term $\mathbb{E}f(\mathbf{x}^{tq}) - \mathbb{E}f(\mathbf{x}^{(t-1)q})$ in the above inequality can be upper-bounded by

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}^{tq}) - \mathbb{E}f(\mathbf{x}^{(t-1)q}) \overset{\text{(i)}}{\leq} & \left(\frac{\eta}{2}-\frac{\eta^2 L}{2}-\frac{3L^2\eta^3 q}{|\mathcal{S}_2|}\right)\sum_{t_1=(t-1)q}^{tq-1}\mathbb{E}\|\mathbf{v}^{t_1}\|^2 + \frac{q^2}{|\mathcal{S}_2|}3\eta L^2 d\delta^2 \\
& + q\eta\pi(\mathcal{S}_1,\delta),
\end{aligned} \tag{82}$$

where (i) is obtained by letting $k_0 = (t-1)$ and $k = tq-1$ in (80). Combining (81) and (82) yields

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}^{K}) \leq & \mathbb{E}f(\mathbf{x}^{0}) - \left(\frac{\eta}{2}-\frac{\eta^2 L}{2}-\frac{3L^2\eta^3 q}{|\mathcal{S}_2|}\right)\sum_{t=0}^{K-1}\mathbb{E}\|\mathbf{v}^{t}\|^2 + \frac{(K-qk^\star)q}{|\mathcal{S}_2|}3\eta L^2 d\delta^2 \\
& + \frac{k^\star q^2}{|\mathcal{S}_2|}3\eta L^2 d\delta^2 + K\eta\pi(\mathcal{S}_1,\delta) \\
\leq & \mathbb{E}f(\mathbf{x}^{0}) - \left(\frac{\eta}{2}-\frac{\eta^2 L}{2}-\frac{3L^2\eta^3 q}{|\mathcal{S}_2|}\right)\sum_{t=0}^{K-1}\mathbb{E}\|\mathbf{v}^{t}\|^2 + \frac{Kq}{|\mathcal{S}_2|}3\eta L^2 d\delta^2 \\
& + K\eta\pi(\mathcal{S}_1,\delta),
\end{aligned} \tag{83}$$

which, in conjunction with (79), yields

$$\mathbb{E}f(\mathbf{x}^K) \leq \mathbb{E}f(\mathbf{x}^0) - \left(\frac{\eta}{2} - \frac{\eta^2 L}{2} - \frac{3L^2\eta^3 q}{|\mathcal{S}_2|}\right)\sum_{t=0}^{K-1}\mathbb{E}\|\mathbf{v}^t\|^2 + \frac{3Kq}{|\mathcal{S}_2|}\eta L^2 d\delta^2 + K\eta\left(\frac{3I(|\mathcal{S}_1|<n)}{|\mathcal{S}_1|}\left(2L^2d\delta^2+\sigma^2\right) + L^2d\delta^2\right). \tag{84}$$

Plugging the notations in Theorem 3 into (84) yields

$$\phi\sum_{t=0}^{K}\mathbb{E}\|\mathbf{v}^t\|^2 \leq \Delta + \frac{3(K+1)q}{|\mathcal{S}_2|}\eta L^2 d\delta^2 + (K+1)\eta\left(\frac{3I(|\mathcal{S}_1|<n)}{|\mathcal{S}_1|}\left(2L^2d\delta^2+\sigma^2\right) + L^2d\delta^2\right), \tag{85}$$

where $\Delta := f(\mathbf{x}^0) - f(\mathbf{x}^*)$ with $\mathbf{x}^* := \arg\min_{\mathbf{x}} f(\mathbf{x})$.

As ζ is generated from $\{0, ..., K\}$ uniformly at random, we have the output $\mathbf{x}^\zeta$ satisfies

$$\begin{aligned}\mathbb{E}\|f(\mathbf{x}^\zeta)\|^2 \leq& 3\mathbb{E}\left(\|f(\mathbf{x}^\zeta) - \hat{\nabla}_{\text{coord}}f(\mathbf{x}^\zeta)\|^2 + \|\mathbf{v}^\zeta - \hat{\nabla}_{\text{coord}}f(\mathbf{x}^\zeta)\|^2 + \|\mathbf{v}^\zeta\|^2\right)\\ \leq& 3L^2d\delta^2 + \underbrace{\frac{3}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\mathbf{v}^k - \hat{\nabla}_{\text{coord}}f(\mathbf{x}^k)\|^2}_{\text{(A)}} + \frac{3}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\mathbf{v}^k\|^2.\end{aligned} \tag{86}$$

We next upper-bound the second term (A) in the above inequality. First note that

$$\frac{1}{K+1}\sum_{t=0}^{K}\mathbb{E}\|\hat{\nabla}_{\text{coord}}f(\mathbf{x}^t) - \mathbf{v}^t\| = \frac{1}{K+1}\left(\sum_{p=0}^{k^\star-1}\sum_{t=pq}^{(p+1)q-1}\mathbb{E}\|\hat{\nabla}_{\text{coord}}f(\mathbf{x}^t) - \mathbf{v}^t\|^2 + \sum_{t=qk^\star}^{K}\mathbb{E}\|\hat{\nabla}_{\text{coord}}f(\mathbf{x}^t) - \mathbf{v}^t\|^2\right).$$

Applying Lemma 6 to the above equation yields

$$\begin{aligned}&\sum_{t=0}^{K}\mathbb{E}\|\hat{\nabla}_{\text{coord}}f(\mathbf{x}^t) - \mathbf{v}^t\| - \sum_{t=qk^\star}^{K}\mathbb{E}\|\hat{\nabla}_{\text{coord}}f(\mathbf{x}^t) - \mathbf{v}^t\|^2\\ &\leq \sum_{p=0}^{k^\star-1}\sum_{t=pq}^{(p+1)q-1}\left(\frac{3L^2\eta^2}{|\mathcal{S}_2|}\sum_{t_1=pq}^{t-1}\mathbb{E}\|\mathbf{v}^{t_1}\|^2 + (t-pq)\frac{6L^2d\delta^2}{|\mathcal{S}_2|} + \pi(\mathcal{S}_1,\delta)\right)\\ &\leq \sum_{p=0}^{k^\star-1}\left(\frac{3qL^2\eta^2}{|\mathcal{S}_2|}\sum_{t_1=pq}^{(p+1)q-1}\mathbb{E}\|\mathbf{v}^{t_1}\|^2 + q(q-1)\frac{3L^2d\delta^2}{|\mathcal{S}_2|} + q\pi(\mathcal{S}_1,\delta)\right)\\ &\leq \frac{3qL^2\eta^2}{|\mathcal{S}_2|}\sum_{p=0}^{k^\star-1}\sum_{t_1=pq}^{(p+1)q-1}\mathbb{E}\|\mathbf{v}^{t_1}\|^2 + qk^\star(q-1)\frac{3L^2d\delta^2}{|\mathcal{S}_2|} + qk^\star\pi(\mathcal{S}_1,\delta),\end{aligned}$$

which, by applying Lemma 6 to $\sum_{t=qk^\star}^{K}\mathbb{E}\|\hat{\nabla}_{\text{coord}}f(\mathbf{x}^t) - \mathbf{v}^t\|^2$, yields

$$\sum_{t=0}^{K}\mathbb{E}\|\hat{\nabla}_{\text{coord}}f(\mathbf{x}^t) - \mathbf{v}^t\| \leq \frac{3qL^2\eta^2}{|\mathcal{S}_2|}\sum_{t=0}^{K}\mathbb{E}\|\mathbf{v}^t\|^2 + \frac{(K-qk^\star)q + q^2k^\star}{|\mathcal{S}_2|}3L^2d\delta^2 + (K+1)\pi(\mathcal{S}_1,\delta). \tag{87}$$

The above inequality can be further simplified to

$$\frac{1}{K+1}\sum_{t=0}^{K}\mathbb{E}\|\hat{\nabla}_{\text{coord}}f(\mathbf{x}^t) - \mathbf{v}^t\| \leq \frac{3qL^2\eta^2}{|\mathcal{S}_2|}\frac{1}{K+1}\sum_{t=0}^{K}\mathbb{E}\|\mathbf{v}^t\|^2 + \frac{3q}{|\mathcal{S}_2|}L^2d\delta^2 + \pi(\mathcal{S}_1,\delta). \tag{88}$$

Combining (85), (86) and (88) yields

$$\mathbb{E}\|f(\mathbf{x}^\zeta)\|^2 \leq 3L^2d\delta^2 + \frac{1}{\phi}\left(\frac{9q\eta^2L^2}{|\mathcal{S}_2|} + 3\right)\left(\frac{\Delta}{K+1} + \eta\left(\theta + L^2d\delta^2\right)\right) + 3\theta,$$

which finishes the proof.

F.3. Proof of Corollary 3

We prove two cases when $n \leq K$ and $n > K$, separately.

First we suppose $n \leq K$. Under the selection of parameters in (14), we have $|\mathcal{S}_1| = n, q = |\mathcal{S}_1| = \lceil n^{1/2} \rceil$, and thus obtain

$$\phi = \eta\left(\frac{1}{2} - \frac{1}{8} - \frac{3}{16}\right) = \frac{3\eta}{16},\ \theta = \frac{3}{K},$$

which, in conjunction with (13), yields

$$\mathbb{E}\|f(\mathbf{x}^\zeta)\|^2 \leq \frac{3}{K} + \left(\frac{16}{3\eta}\frac{9}{16} + \frac{16}{\eta}\right)\left(\frac{\Delta}{K} + \eta\left(\frac{4}{K}\right)\right) + \frac{9}{K} = \frac{76\Delta L + 88}{K} \leq \mathcal{O}\left(\frac{1}{K}\right).$$

We choose $K = C\epsilon^{-1}$, where $C > 0$ is a constant. Then, based on the above inequality, we have, for C large enough, our Algorithm 2 achieves $\mathbb{E}\|f(\mathbf{x}^\zeta)\|^2 \leq \epsilon$, and the total number of function queries can be bounded as

$$\left\lceil \frac{K}{q} \right\rceil nd + K|\mathcal{S}_2|d \leq Kn^{1/2}d + nd + Kn^{1/2}d + Kd \leq \mathcal{O}(nd + \epsilon^{-1}n^{1/2}d) \leq \mathcal{O}(\epsilon^{-1}n^{1/2}d) \leq \mathcal{O}(d\epsilon^{-3/2}), \tag{89}$$

where the last two inequalities follow from the assumption that $n \leq K = C\epsilon^{-1}$.

Next, we suppose $n > K$. In this case, we have $|\mathcal{S}_1| = K, q = |\mathcal{S}_1| = \lceil K^{1/2} \rceil$, and

$$\phi = \eta\left(\frac{1}{2} - \frac{1}{8} - \frac{3}{16}\right) = \frac{3\eta}{16},\ \theta = \frac{3}{K} + \frac{3}{K}\left(\frac{2}{K} + \sigma^2\right),$$

which, in conjunction with (13), yields

$$\mathbb{E}\|f(\mathbf{x}^\zeta)\|^2 \leq \frac{3}{K} + \frac{19}{\eta}\left(\frac{\Delta}{K} + \eta\left(\frac{4+3\sigma^2}{K} + \frac{6}{K^2}\right)\right) + \frac{9}{K} = \frac{76\Delta L + 88 + 57\sigma^2}{K} + \frac{114}{K^2} \leq \mathcal{O}\left(\frac{1}{K}\right).$$

We choose $K = C\epsilon^{-1}$, where $C > 0$ is a constant. Then, for C large enough, our Algorithm 2 achieves $\mathbb{E}\|f(\mathbf{x}^\zeta)\|^2 \leq \epsilon$, and the total number of function queries can be bounded as

$$\begin{aligned}\left\lceil \frac{K}{q} \right\rceil Kd + K|\mathcal{S}_2|d &\leq Kd + K^{3/2}d + K^{3/2}d + Kd = 2K^{3/2}d + 2Kd \leq \mathcal{O}(K^{3/2}d) \\ &\leq \mathcal{O}(d\epsilon^{-3/2}) \leq \mathcal{O}(\epsilon^{-1}n^{1/2}d),\end{aligned} \tag{90}$$

where the last inequality follows from the assumption that $n > K \geq C\epsilon^{-1}$.

Combining (89) and (90) implies that the number of function queries required by Algorithm 2 is at most $\mathcal{O}\left(\min\{\epsilon^{-1}n^{1/2}d, d\epsilon^{-3/2}\}\right)$.

F.4. Proof of Corollary 4

We prove two cases when $n \leq \lceil K^{2/3} \rceil$ and $n > \lceil K^{2/3} \rceil$, separately.

First we suppose $n \leq \lceil K^{2/3} \rceil$, and thus we have $q = |\mathcal{S}_1| = n, \eta = 1/(4L\sqrt{n}), \delta = 1/(L\sqrt{nKd})$. Then, we obtain

$$\phi = \eta\left(\frac{1}{2} - \frac{1}{8\sqrt{n}} - \frac{3}{16}\right) \geq \frac{3\eta}{16},\ \theta = \frac{3}{K},$$

which, in conjunction with (13), yields

$$\mathbb{E}\|f(\mathbf{x}^\zeta)\|^2 \leq \frac{76\sqrt{n}L\delta}{K} + \frac{66}{K} + \frac{22}{nK} = \frac{76\sqrt{|\mathcal{S}_1|}L\delta}{K} + \frac{66}{K} + \frac{22}{|\mathcal{S}_1|K} \leq \mathcal{O}\left(\frac{\sqrt{|\mathcal{S}_1|}}{K}\right).$$

Let $K = C\sqrt{n}\epsilon^{-1}$ for a positive constant C, which, combined with $n \leq \lceil K^{2/3} \rceil$, implies that $n \leq \mathcal{O}(\epsilon^{-1})$. Then, our Algorithm 2 achieves $\mathbb{E}\|f(\mathbf{x}^\zeta)\|^2 \leq \epsilon$, and the total number of function queries can be bounded as

$$\left\lceil \frac{K}{q} \right\rceil nd + Kd \leq nd + 2Kd \leq \mathcal{O}\left(nd + \frac{n^{1/2}d}{\epsilon}\right) \leq \mathcal{O}(\epsilon^{-1}n^{1/2}d) \leq \mathcal{O}\left(\frac{d}{\epsilon^{3/2}}\right), \tag{91}$$

where the last two inequalities follow from the assumption that $n \leq \mathcal{O}(\epsilon^{-1})$.

Next, we suppose $n > \lceil K^{2/3} \rceil$. In this case, we have $|\mathcal{S}_1| = \lceil K^{2/3} \rceil$, and thus

$$\phi \geq \eta\left(\frac{1}{2} - \frac{1}{8} - \frac{3}{16}\right) \geq \frac{3\eta}{16},\ \theta = \frac{3}{K} + \frac{6}{K^{7/3}} + \frac{3\sigma^2}{K^{2/3}}, \tag{92}$$

which, in conjunction with (13), yields

$$\mathbb{E}\|f(\mathbf{x}^\zeta)\|^2 \leq \mathcal{O}\left(\frac{K^{1/3}}{K}\right) \leq \mathcal{O}\left(\frac{\sqrt{|\mathcal{S}_1|}}{K}\right),$$

where the last inequality follows from the assumption that $\lceil K^{2/3} \rceil < n$. Let $K = C\epsilon^{-3/2}$, where $C > 0$ is a constant. Then, for C large enough, our Algorithm 2 achieves $\mathbb{E}\|f(\mathbf{x}^\zeta)\|^2 \leq \epsilon$, and the total number of function queries can be bounded as

$$\left\lceil \frac{K}{q} \right\rceil |\mathcal{S}_1|d + Kd \leq 2Kd + |\mathcal{S}_1|d \leq 3Kd \leq \mathcal{O}\left(\frac{d}{\epsilon^{3/2}}\right) \leq \mathcal{O}\left(\frac{n^{1/2}d}{\epsilon}\right) \tag{93}$$

where the last inequality follows from the assumption that $n > \lceil K^{2/3} \rceil = C^{2/3}\epsilon^{-1}$.

Combining the above two cases finishes the proof.

G. Proof for ZO-SPIDER-Coord under PL Condition

G.1. Proof of Theorem 4

Let $\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$. Then, for any $qk_0 \leq m \leq q(k_0+1) - 1, k_0 = 0, ..., h-1$ $(h = K/q)$, we have

$$\begin{aligned}
&f(\mathbf{x}^{m+1}) \\
&\leq f(\mathbf{x}^m) + \langle \nabla f(\mathbf{x}^m), \mathbf{x}^{m+1} - \mathbf{x}^m \rangle + \frac{L}{2}\|\mathbf{x}^{m+1} - \mathbf{x}^m\|^2 \\
&= f(\mathbf{x}^m) - \frac{\eta}{2}\langle \nabla f(\mathbf{x}^m) - \mathbf{v}^m, \mathbf{v}^m \rangle - \frac{\eta}{2}\|\mathbf{v}^m\|^2 - \frac{\eta}{2}\langle \nabla f(\mathbf{x}^m), \mathbf{v}^m - \nabla f(\mathbf{x}^m) \rangle - \frac{\eta}{2}\|\nabla f(\mathbf{x}^m)\|^2 + \frac{L\eta^2}{2}\|\mathbf{v}^m\|^2 \\
&\leq f(\mathbf{x}^m) + \frac{\eta}{2}\|\nabla f(\mathbf{x}^m) - \mathbf{v}^m\|^2 - \frac{\eta}{4}\|\mathbf{v}^m\|^2 - \frac{\eta}{4}\|\nabla f(\mathbf{x}^m)\|^2 + \frac{L\eta^2}{2}\|\mathbf{v}^m\|^2 \\
&= f(\mathbf{x}^m) + \frac{\eta}{2}\|\nabla f(\mathbf{x}^m) - \mathbf{v}^m\|^2 - \Big(\frac{\eta}{4} - \frac{L\eta^2}{2}\Big)\|\mathbf{v}^m\|^2 - \frac{\eta}{4}\|\nabla f(\mathbf{x}^m)\|^2, \\
&\overset{\text{(i)}}{\leq} f(\mathbf{x}^m) + \frac{\eta}{2}\|\nabla f(\mathbf{x}^m) - \mathbf{v}^m\|^2 - \Big(\frac{\eta}{4} - \frac{L\eta^2}{2}\Big)\|\mathbf{v}^m\|^2 - \frac{\eta}{4\gamma}(f(\mathbf{x}^m) - f(\mathbf{x}^*))
\end{aligned} \tag{94}$$

where (i) follows from Definition 1. Taking expectation over the above inequality and using Lemma 6, we have

$$\begin{aligned}
\mathbb{E}(f(\mathbf{x}^{m+1}) - f(\mathbf{x}^*)) \leq &\left(1 - \frac{\eta}{4\gamma}\right)\mathbb{E}(f(\mathbf{x}^m) - f(\mathbf{x}^*)) + \frac{3\eta^3 L^2}{|\mathcal{S}_2|}\sum_{t=qk_0}^{m}\mathbb{E}\|\mathbf{v}^t\|^2 + (m - qk_0)\frac{6\eta L^2 d\delta^2}{|\mathcal{S}_2|} + \eta L^2 d\delta^2 \\
&- \Big(\frac{\eta}{4} - \frac{L\eta^2}{2}\Big)\mathbb{E}\|\mathbf{v}^m\|^2.
\end{aligned} \tag{95}$$

To simplify notation, we let $\alpha := 1 - \eta/(4\gamma)$. Then, telescoping (95) over m from qk_0 to $q(k_0+1)-1$ yields

$$\begin{aligned}
\mathbb{E}(f(\mathbf{x}^{q(k_0+1)}) - f(\mathbf{x}^*)) \leq{}& \alpha^q \mathbb{E}(f(\mathbf{x}^{qk_0}) - f(\mathbf{x}^*)) + \sum_{m=qk_0}^{q(k_0+1)-1} \frac{3\eta^3 L^2}{|\mathcal{S}_2|} \alpha^{q(k_0+1)-m-1} \sum_{t=qk_0}^{m} \mathbb{E}\|\mathbf{v}^t\|^2 \\
&+ \sum_{m=qk_0}^{q(k_0+1)-1} \left(\frac{b_\gamma}{B_\gamma} + 1\right) \alpha^{q(k_0+1)-m-1} \eta L^2 d\delta^2 - \left(\frac{\eta}{4} - \frac{L\eta^2}{2}\right) \sum_{m=qk_0}^{q(k_0+1)-1} \alpha^{q(k_0+1)-m-1} \mathbb{E}\|\mathbf{v}^m\|^2 \\
={}& \alpha^q \mathbb{E}(f(\mathbf{x}^{qk_0}) - f(\mathbf{x}^*)) + \frac{3\eta^3 L^2}{|\mathcal{S}_2|} \sum_{m=qk_0}^{q(k_0+1)-1} \sum_{t=m}^{q(k_0+1)-1} \alpha^{q(k_0+1)-t-1} \mathbb{E}\|\mathbf{v}^m\|^2 \\
&+ \sum_{m=qk_0}^{q(k_0+1)-1} \left(\frac{b_\gamma}{B_\gamma} + 1\right) \alpha^{q(k_0+1)-m-1} \eta L^2 d\delta^2 - \left(\frac{\eta}{4} - \frac{L\eta^2}{2}\right) \sum_{m=qk_0}^{q(k_0+1)-1} \alpha^{q(k_0+1)-m-1} \mathbb{E}\|\mathbf{v}^m\|^2
\end{aligned}$$

where the first inequality follows from the fact that $(m-qk_0)/|\mathcal{S}_2| \leq q/(B_\gamma \gamma L) \leq b_\gamma/B_\gamma$. Noting that $\alpha^{q(k_0+1)-m-1} \geq \alpha^q$ and $\sum_{t=m}^{q(k_0+1)-1} \alpha^{q(k_0+1)-t-1} = (1-\alpha^{q(k_0+1)-m})/(1-\alpha) < 1/(1-\alpha) = 4\gamma/\eta$, we obtain from the above inequality that

$$\begin{aligned}
&\mathbb{E}(f(\mathbf{x}^{q(k_0+1)}) - f(\mathbf{x}^*)) \\
&\leq \alpha^q \mathbb{E}(f(\mathbf{x}^{qk_0}) - f(\mathbf{x}^*)) + \frac{12\gamma\eta^2 L^2}{|\mathcal{S}_2|} \sum_{m=qk_0}^{q(k_0+1)-1} \mathbb{E}\|\mathbf{v}^m\|^2 \\
&\quad + \frac{1-\alpha^q}{1-\alpha} \left(\frac{b_\gamma}{B_\gamma} + 1\right) \eta L^2 d\delta^2 - \left(\frac{\eta}{4} - \frac{L\eta^2}{2}\right) \sum_{m=qk_0}^{q(k_0+1)-1} \alpha^q \mathbb{E}\|\mathbf{v}^m\|^2 \\
&\overset{\text{(i)}}{\leq} \alpha^q \mathbb{E}(f(\mathbf{x}^{qk_0}) - f(\mathbf{x}^*)) + \frac{12\eta^2 L}{B_\gamma} \sum_{m=qk_0}^{q(k_0+1)-1} \mathbb{E}\|\mathbf{v}^m\|^2 \\
&\quad + \frac{1-\alpha^q}{1-\alpha} \left(\frac{b_\gamma}{B_\gamma} + 1\right) \eta L^2 d\delta^2 - \left(\frac{\eta}{4} - \frac{L\eta^2}{2}\right) \left(1 - \frac{b_\gamma}{16q}\right)^q \sum_{m=qk_0}^{q(k_0+1)-1} \mathbb{E}\|\mathbf{v}^m\|^2 \\
&= \alpha^q \mathbb{E}(f(\mathbf{x}^{qk_0}) - f(\mathbf{x}^*)) + \frac{1-\alpha^q}{1-\alpha} \left(\frac{b_\gamma}{B_\gamma} + 1\right) \eta L^2 d\delta^2 - \eta \left(\frac{1}{8}\left(1 - \frac{b_\gamma}{16q}\right)^q - \frac{3}{B_\gamma}\right) \sum_{m=qk_0}^{q(k_0+1)-1} \mathbb{E}\|\mathbf{v}^m\|^2 \\
&\leq \alpha^q \mathbb{E}(f(\mathbf{x}^{qk_0}) - f(\mathbf{x}^*)) + \frac{1-\alpha^q}{1-\alpha} \left(\frac{b_\gamma}{B_\gamma} + 1\right) \eta L^2 d\delta^2
\end{aligned} \tag{96}$$

where (i) follows from the facts that $\frac{q}{b_\gamma L} < \gamma$ and $|\mathcal{S}_2| = \lceil \gamma L B_\gamma \rceil$ and the last inequality follows from the condition that $\frac{1}{8}\left(1 - \frac{b_\gamma}{16q}\right)^q - \frac{3}{B_\gamma} > 0$.

Telescoping (96) over k_0 from 0 to $h-1$ yields

$$\begin{aligned}
\mathbb{E}(f(\mathbf{x}^K) - f(\mathbf{x}^*)) &\leq \left(1 - \frac{\eta}{4\gamma}\right)^K (f(\mathbf{x}^0) - f(\mathbf{x}^*)) + \sum_{k_0=0}^{h-1} \alpha^{qk_0} \frac{1-\alpha^q}{1-\alpha} \left(\frac{b_\gamma}{B_\gamma} + 1\right) \eta L^2 d\delta^2 \\
&= \left(1 - \frac{\eta}{4\gamma}\right)^K (f(\mathbf{x}^0) - f(\mathbf{x}^*)) + \frac{1-\alpha^k}{1-\alpha^q} \frac{1-\alpha^q}{1-\alpha} \left(\frac{b_\gamma}{B_\gamma} + 1\right) \eta L^2 d\delta^2 \\
&\leq \left(1 - \frac{1}{16L\gamma}\right)^K (f(\mathbf{x}^0) - f(\mathbf{x}^*)) + 4\gamma \left(\frac{b_\gamma}{B_\gamma} + 1\right) L^2 d\delta^2.
\end{aligned} \tag{97}$$

From (97), we require the total number $K = \mathcal{O}(\gamma \log(1/\epsilon))$ and $\delta = \mathcal{O}(\sqrt{\epsilon}/(L\sqrt{\gamma d}))$ to achieve $\mathbb{E}(f(\mathbf{x}^K) - f(\mathbf{x}^*)) < \epsilon$. Thus, the total number of function queries is $\left\lceil \frac{K}{q} \right\rceil nd + K|\mathcal{S}_2|d = \mathcal{O}\left(d(\gamma n^{1/2} + \gamma^2) \log\left(\frac{1}{\epsilon}\right)\right)$. Then, the proof is complete.

H. Proofs for PROX-ZO-SPIDER-Coord

H.1. Auxiliary Lemma

We first prove the following useful lemma.

Lemma 7. *Let Assumption 1 hold, and define*

$$\tau = \frac{\eta}{2} - \frac{L\eta^2}{2} - \frac{3q}{|\mathcal{S}_2|}\eta^3 L^2$$
$$C = \frac{6I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right) + 2L^2 d\delta^2. \tag{98}$$

Then, we have,

$$\mathbb{E}\|G(\mathbf{x}^\zeta, \nabla f(\mathbf{x}^\zeta), \eta)\|^2 \leq \frac{2}{\tau}\left(\frac{\Delta_\psi}{K} + \frac{3q}{|\mathcal{S}_2|}\eta L^2 d\delta^2 + \frac{\eta}{2}C\right) + \frac{12qL^2\eta^2}{|\mathcal{S}_2|\tau}\left(\frac{\Delta_\psi}{K} + \frac{3q}{|\mathcal{S}_2|}\eta L^2 d\delta^2 + \frac{\eta}{2}C\right) + \frac{12q}{|\mathcal{S}_2|}L^2 d\delta^2 + 4C, \tag{99}$$

where $\Delta_\psi = \psi(\mathbf{x}^0) - \psi(\mathbf{x}^*)$ *with* $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d}\psi(\mathbf{x})$.

Proof. We first introduce the following notation for our proof

$$G(\mathbf{x}, \mathbf{g}, \eta) = \frac{1}{\eta}(\mathbf{x} - \mathbf{x}_g), \text{ where } \mathbf{x}_g = \arg\min_{\mathbf{z}\in\mathbb{R}^d}\left\{\langle \mathbf{g}, \mathbf{z}\rangle + \frac{1}{2\eta}\|\mathbf{z} - \mathbf{x}\|^2 + h(\mathbf{z})\right\}. \tag{100}$$

Note that when $\mathbf{g} = \nabla f(\mathbf{x})$, $G(\mathbf{x}, \mathbf{g}, \eta)$ becomes the generalized projected gradient of the objective $\Psi(\cdot)$ at $\mathbf{x}$. The following lemma provides important properties of $G(\mathbf{x}, \mathbf{g}, \eta)$ by Lemma 1 and Proposition 1 in Ghadimi et al. 2016.

Lemma 8. *For any* $\mathbf{g}, \mathbf{g}_1$ *and* $\mathbf{g}_2$ *in* $\mathbb{R}^d$*, we have*

(i) $\langle \mathbf{g}, G(\mathbf{x}, \mathbf{g}, \eta)\rangle \geq \|G(\mathbf{x}, \mathbf{g}, \eta)\|^2 + (h(\mathbf{x}_g) - h(\mathbf{x}))/\eta$, *where* $\mathbf{x}_g$ *is defined by* (100).

(ii) $\|G(\mathbf{x}, \mathbf{g}_1, \eta) - G(\mathbf{x}, \mathbf{g}_2, \eta)\| \leq \|\mathbf{g}_1 - \mathbf{g}_2\|$.

Based on the above results, we now prove Lemma 7. Using an approach similar to Lemma 6, we obtain, for any given $k_0 \leq \lfloor K/q\rfloor$ and $qk_0 \leq k \leq \min\{q(k_0+1)-1, K\}$,

$$\mathbb{E}\|\mathbf{v}^k - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)\|^2 \leq \frac{3L^2}{|\mathcal{S}_2|}\sum_{t=qk_0}^{k-1}\mathbb{E}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 + (k - qk_0)\frac{6L^2 d\delta^2}{|\mathcal{S}_2|} + \frac{3I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right),$$

which, based on the proximal gradient step and (100), implies that $\mathbf{x}^k - \mathbf{x}^{k+1} = \eta G(\mathbf{x}^k, \mathbf{v}^k, \eta)$. Thus

$$\mathbb{E}\|\mathbf{v}^k - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)\|^2 \leq \frac{3L^2\eta^2}{|\mathcal{S}_2|}\sum_{t=qk_0}^{k-1}\mathbb{E}\|G(\mathbf{x}^t, \mathbf{v}^t, \eta)\|^2 + (k - qk_0)\frac{6L^2 d\delta^2}{|\mathcal{S}_2|} + \frac{3I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right),$$

which, in conjunction with (3), implies that

$$\begin{aligned}\mathbb{E}\|\mathbf{v}^k - \nabla f(\mathbf{x}^k)\|^2 &\leq 2\mathbb{E}\|\mathbf{v}^k - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)\|^2 + 2\mathbb{E}\|\nabla f(\mathbf{x}^k) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^k)\|^2 \\ &\leq \frac{6L^2\eta^2}{|\mathcal{S}_2|}\sum_{t=qk_0}^{k-1}\mathbb{E}\|G(\mathbf{x}^t, \mathbf{v}^t, \eta)\|^2 + (k - qk_0)\frac{12L^2 d\delta^2}{|\mathcal{S}_2|} \\ &\quad + \frac{6I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right) + 2L^2 d\delta^2. \end{aligned} \tag{101}$$

Recalling that the gradient $\nabla f(\mathbf{x})$ is L-Lipschitz, we have, for any $k_0 \leq \lfloor K/q \rfloor$ and $qk_0 \leq k \leq \min\{q(k_0+1)-1, K\}$,

$$\begin{aligned}
f(\mathbf{x}^{k+1}) \leq & f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \\
= & f(\mathbf{x}^k) - \eta \langle \nabla f(\mathbf{x}^k), G(\mathbf{x}^k, \mathbf{v}^k, \eta) \rangle + \frac{L\eta^2}{2}\|G(\mathbf{x}^k, \mathbf{v}^k, \eta)\|^2 \\
= & f(\mathbf{x}^k) - \eta \langle \nabla f(\mathbf{x}^k) - \mathbf{v}^k, G(\mathbf{x}^k, \mathbf{v}^k, \eta) \rangle - \eta \langle \mathbf{v}^k, G(\mathbf{x}^k, \mathbf{v}^k, \eta) \rangle + \frac{L\eta^2}{2}\|G(\mathbf{x}^k, \mathbf{v}^k, \eta)\|^2 \\
\leq & f(\mathbf{x}^k) + \frac{\eta}{2}\|\nabla f(\mathbf{x}^k) - \mathbf{v}^k\|^2 + \frac{\eta}{2}\|G(\mathbf{x}^k, \mathbf{v}^k, \eta)\|^2 - \eta \langle \mathbf{v}^k, G(\mathbf{x}^k, \mathbf{v}^k, \eta) \rangle + \frac{L\eta^2}{2}\|G(\mathbf{x}^k, \mathbf{v}^k, \eta)\|^2,
\end{aligned}$$

which, in conjunction with Lemma 8, implies that

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \frac{\eta}{2}\|\nabla f(\mathbf{x}^k) - \mathbf{v}^k\|^2 - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\|G(\mathbf{x}^k, \mathbf{v}^k, \eta)\|^2 - (h(\mathbf{x}^{k+1}) - h(\mathbf{x}^k)). \tag{102}$$

Let $\psi(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x})$. Then, taking expectation over (102) yields

$$\mathbb{E}\psi(\mathbf{x}^{k+1}) \leq \mathbb{E}\psi(\mathbf{x}^k) + \frac{\eta}{2}\mathbb{E}\|\nabla f(\mathbf{x}^k) - \mathbf{v}^k\|^2 - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\mathbb{E}\|G(\mathbf{x}^k, \mathbf{v}^k, \eta)\|^2.$$

Telescoping the above inequality yields, for $qk_0 \leq k \leq \min\{q(k_0+1)-1, K\}$,

$$\mathbb{E}\psi(\mathbf{x}^{k+1}) \leq \mathbb{E}\psi(\mathbf{x}^{k_0 q}) + \frac{\eta}{2}\sum_{t=k_0 q}^{k}\mathbb{E}\|\nabla f(\mathbf{x}^t) - \mathbf{v}^t\|^2 - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\sum_{t=k_0 q}^{k}\mathbb{E}\|G(\mathbf{x}^t, \mathbf{v}^t, \eta)\|^2,$$

which, recalling the definition of C in (98) and using (101), implies that

$$\begin{aligned}
\mathbb{E}\psi(\mathbf{x}^{k+1}) \leq & \mathbb{E}\psi(\mathbf{x}^{k_0 q}) + \frac{\eta}{2}\sum_{t=k_0 q}^{k}\left(\frac{6L^2\eta^2}{|\mathcal{S}_2|}\sum_{p=qk_0}^{t-1}\mathbb{E}\|G(\mathbf{x}^p, \mathbf{v}^p, \eta)\|^2 + (t - qk_0)\frac{12L^2 d\delta^2}{|\mathcal{S}_2|} + C\right) \\
& - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\sum_{t=k_0 q}^{k}\|G(\mathbf{x}^t, \mathbf{v}^t, \eta)\|^2 \\
\leq & \mathbb{E}\psi(\mathbf{x}^{k_0 q}) + \frac{3L^2\eta^3}{|\mathcal{S}_2|}\sum_{t=k_0 q}^{k}\sum_{p=qk_0}^{t-1}\mathbb{E}\|G(\mathbf{x}^p, \mathbf{v}^p, \eta)\|^2 + \frac{\eta}{2}\sum_{t=k_0 q}^{k}(t - qk_0)\frac{12L^2 d\delta^2}{|\mathcal{S}_2|} \\
& + \frac{\eta}{2}\sum_{t=k_0 q}^{k} C - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\sum_{t=k_0 q}^{k}\|G(\mathbf{x}^t, \mathbf{v}^t, \eta)\|^2 \\
\leq & \mathbb{E}\psi(\mathbf{x}^{k_0 q}) + \frac{3L^2\eta^3}{|\mathcal{S}_2|}(k - k_0 q + 1)\sum_{p=qk_0}^{k-1}\mathbb{E}\|G(\mathbf{x}^p, \mathbf{v}^p, \eta)\|^2 + \frac{3(k - qk_0 + 1)(k - qk_0)}{|\mathcal{S}_2|}\eta L^2 d\sigma^2 \\
& + \frac{\eta}{2}(k - qk_0 + 1)C - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\sum_{t=k_0 q}^{k}\|G(\mathbf{x}^t, \mathbf{v}^t, \eta)\|^2.
\end{aligned} \tag{103}$$

Using the above inequality and letting $K^* = \lfloor K/q \rfloor$, we obtain

$$\begin{aligned}
\mathbb{E}\psi(\mathbf{x}^K) - \mathbb{E}\psi(\mathbf{x}^0) \leq & \mathbb{E}\psi(\mathbf{x}^K) - \mathbb{E}\psi(\mathbf{x}^{qK^*}) + \sum_{i=1}^{K^*} \Big(\mathbb{E}\psi(\mathbf{x}^{qi}) - \mathbb{E}\psi(\mathbf{x}^{q(i-1)}) \Big) \\
\leq & \frac{3L^2\eta^3}{|\mathcal{S}_2|}(K - qK^*) \sum_{p=qK^*}^{K-2} \mathbb{E}\|G(\mathbf{x}^p, \mathbf{v}^p, \eta)\|^2 + \frac{3(K-qK^*)(K-qK^*-1)}{|\mathcal{S}_2|}\eta L^2 d\delta^2 \\
& + \frac{\eta}{2}(K - qK^*)C - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right) \sum_{t=qK^*}^{K-1} \|G(\mathbf{x}^t, \mathbf{v}^t, \eta)\|^2 \\
& + \sum_{i=1}^{K^*} \Bigg(\frac{3\eta^3 L^2 q}{|\mathcal{S}_2|} \sum_{p=q(i-1)}^{qi-1} \mathbb{E}\|G(\mathbf{x}^p, \mathbf{v}^p, \eta)\|^2 + \frac{3q^2\eta L^2 d\delta^2}{|\mathcal{S}_2|} + \frac{q\eta}{2}C \\
& - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right) \sum_{t=q(i-1)}^{qi-1} \|G(\mathbf{x}^t, \mathbf{v}^t, \eta)\|^2 \Bigg) \\
\leq & \frac{3L^2\eta^3}{|\mathcal{S}_2|}(K - qK^*) \sum_{p=qK^*}^{K-2} \mathbb{E}\|G(\mathbf{x}^p, \mathbf{v}^p, \eta)\|^2 + \frac{3(K-qK^*)q}{|\mathcal{S}_2|}\eta L^2 d\delta^2 \\
& + \frac{\eta}{2}(K - qK^*)C - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right) \sum_{t=qK^*}^{K-1} \|G(\mathbf{x}^t, \mathbf{v}^t, \eta)\|^2 \\
& + \frac{3\eta^3 L^2 q}{|\mathcal{S}_2|} \sum_{p=0}^{qK^*-1} \mathbb{E}\|G(\mathbf{x}^p, \mathbf{v}^p, \eta)\|^2 + \frac{3K^*q^2\eta L^2 d\delta^2}{|\mathcal{S}_2|} + \frac{K^*q\eta}{2}C \\
& - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right) \sum_{t=0}^{qK^*-1} \|G(\mathbf{x}^t, \mathbf{v}^t, \eta)\|^2 \\
\leq & - \left(\frac{\eta}{2} - \frac{L\eta^2}{2} - \frac{3\eta^3 L^2 q}{|\mathcal{S}_2|}\right) \sum_{t=0}^{K-1} \mathbb{E}\|G(\mathbf{x}^t, \mathbf{v}^t, \eta)\|^2 + \frac{3Kq}{|\mathcal{S}_2|}\eta L^2 d\delta^2 + \frac{K\eta}{2}C. \qquad (104)
\end{aligned}$$

Based on the above inequality, we have

$$\left(\frac{\eta}{2} - \frac{L\eta^2}{2} - \frac{3\eta^3 L^2 q}{|\mathcal{S}_2|}\right) \sum_{t=0}^{K-1} \mathbb{E}\|G(\mathbf{x}^t, \mathbf{v}^t, \eta)\|^2 \leq \Delta_\psi + \frac{3Kq}{|\mathcal{S}_2|}\eta L^2 d\delta^2 + \frac{K\eta}{2}C, \qquad (105)$$

where $\Delta_\psi = \psi(\mathbf{x}^0) - \psi(\mathbf{x}^*)$ with $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \psi(\mathbf{x})$. To simplify notation, we define

$$\tau = \frac{\eta}{2} - \frac{L\eta^2}{2} - \frac{3q}{|\mathcal{S}_2|}\eta^3 L^2.$$

Then, (105) is simplified to

$$\frac{1}{K}\sum_{t=0}^{K-1} \mathbb{E}\|G(\mathbf{x}^t, \mathbf{v}^t, \eta)\|^2 \leq \frac{\Delta_\psi}{K\tau} + \frac{1}{\tau}\left(\frac{3q}{|\mathcal{S}_2|}\eta L^2 d\delta^2 + \frac{\eta}{2}C\right). \qquad (106)$$

Using the inequality that $\|\mathbf{x} + \mathbf{y}\|^2 \leq 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$, we have

$$\mathbb{E}\|G(\mathbf{x}^\zeta, \nabla f(\mathbf{x}^\zeta), \eta)\|^2 \leq 2\mathbb{E}\|G(\mathbf{x}^\zeta, \mathbf{v}^\zeta, \eta)\|^2 + 2\mathbb{E}\|G(\mathbf{x}^\zeta, \mathbf{v}^\zeta, \eta) - G(\mathbf{x}^\zeta, \nabla f(\mathbf{x}^\zeta), \eta)\|^2. \qquad (107)$$

To upper-bound the first term of the right side of (107), we have

$$\mathbb{E}\|G(\mathbf{x}^\zeta, \mathbf{v}^\zeta, \eta)\|^2 = \frac{1}{K}\sum_{t=1}^{K} \mathbb{E}\|G(\mathbf{x}^t, \mathbf{v}^t, \eta)\|^2 \leq \frac{\Delta_\psi}{K\tau} + \frac{1}{\tau}\left(\frac{3q}{|\mathcal{S}_2|}\eta L^2 d\delta^2 + \frac{\eta}{2}C\right). \qquad (108)$$

For the second term of (107), we have

$$\begin{aligned}\mathbb{E}\|G(\mathbf{x}^\zeta,\mathbf{v}^\zeta,\eta)-G(\mathbf{x}^\zeta,\nabla f(\mathbf{x}^\zeta),\eta)\|^2 \overset{\text{(i)}}{\leq}&\mathbb{E}\|\mathbf{v}^\zeta-\nabla f(\mathbf{x}^\zeta)\|^2 \leq \frac{1}{K}\sum_{t=1}^{K}\mathbb{E}\|\mathbf{v}^t-\nabla f(\mathbf{x}^t)\|^2\\ \overset{\text{(ii)}}{\leq}&\frac{6qL^2\eta^2}{|\mathcal{S}_2|}\frac{1}{K}\sum_{t=0}^{K-1}\mathbb{E}\|G(\mathbf{x}^t,\mathbf{v}^t,\eta)\|^2+\frac{6q}{|\mathcal{S}_2|}L^2d\delta^2+C\\ \leq&\frac{6qL^2\eta^2}{|\mathcal{S}_2|\tau}\left(\frac{\Delta_\psi}{K}+\frac{3q}{|\mathcal{S}_2|}\eta L^2d\delta^2+\frac{\eta}{2}C\right)\\ &+\frac{6q}{|\mathcal{S}_2|}L^2d\delta^2+2C \end{aligned} \tag{109}$$

where (i) follows from Lemma 8, (ii) follows from (101) and the last inequality following from (106). Combining (107), (108) and (109) yields

$$\begin{aligned}\mathbb{E}\|G(\mathbf{x}^\zeta,\nabla f(\mathbf{x}^\zeta),\eta)\|^2 \leq&\frac{2}{\tau}\left(\frac{\Delta_\psi}{K}+\frac{3q}{|\mathcal{S}_2|}\eta L^2d\delta^2+\frac{\eta}{2}C\right)\\ &+\frac{12qL^2\eta^2}{|\mathcal{S}_2|\tau}\left(\frac{\Delta_\psi}{K}+\frac{3q}{|\mathcal{S}_2|}\eta L^2d\delta^2+\frac{\eta}{2}C\right)\\ &+\frac{12q}{|\mathcal{S}_2|}L^2d\delta^2+4C, \end{aligned} \tag{110}$$

which finishes the proof. □

H.2. Proof of Theorem 5

Based on Lemma 7, we next prove our Theorem 5. We prove two cases with $n \leq K$ and $n > K$, respectively.

First we suppose $n \leq K$. Based on the selected parameters, we have $|\mathcal{S}_1| = n, q = |\mathcal{S}_1| = \lceil n^{1/2} \rceil$, and thus obtain

$$\tau = \frac{3\eta}{16},\ C = \frac{2}{K},\ \frac{qL^2\eta^2}{|\mathcal{S}_2|} = \frac{1}{16}$$

which, in conjunction with (99) in Lemma 7, implies that

$$\mathbb{E}\|G(\mathbf{x}^\zeta,\nabla f(\mathbf{x}^\zeta),\eta)\|^2 \leq \frac{32}{3\eta}\left(\frac{\Delta_\psi}{K}+\frac{4\eta}{K}\right)+\frac{4}{\eta}\left(\frac{\Delta_\psi}{K}+\frac{4\eta}{K}\right)+\frac{20}{K} \leq \frac{60\Delta_\psi+80}{K} \leq \mathcal{O}\left(\frac{1}{K}\right).$$

We choose $K = C\epsilon^{-1}$, where C is a positive constant. Then, based on the above inequality, for C large enough, our PROX-ZO-SPIDER-Coord achieves an ϵ-approximate stationary point, i.e., $\mathbb{E}\|G(\mathbf{x}^\zeta,\nabla f(\mathbf{x}^\zeta),\eta)\|^2 \leq \epsilon$, and the total number of function queries is

$$\begin{aligned}\left\lceil\frac{K}{q}\right\rceil nd + K|\mathcal{S}_2|d &\leq Kn^{1/2}d + nd + Kn^{1/2}d + Kd\\ &\leq \mathcal{O}(nd+\epsilon^{-1}n^{1/2}d) \leq\ \mathcal{O}(\epsilon^{-1}n^{1/2}d) \leq \mathcal{O}(d\epsilon^{-3/2}), \end{aligned} \tag{111}$$

where the last two inequalities follow from the assumption that $n \leq K = C\epsilon^{-1}$.

Next, we suppose $n > K$. In this case, we have $|\mathcal{S}_1| = K, q = |\mathcal{S}_1| = \lceil K^{1/2} \rceil$, and

$$\tau = \frac{3\eta}{16},\ C = \frac{6\sigma^2+2}{K}+\frac{12}{K^2},\ \frac{qL^2\eta^2}{|\mathcal{S}_2|} = \frac{1}{16}, \tag{112}$$

which, in conjunction with (99), implies that

$$\begin{aligned}\mathbb{E}\|G(\mathbf{x}^\zeta,\nabla f(\mathbf{x}^\zeta),\eta)\|^2 \leq& \frac{32}{3\eta}\left(\frac{\Delta_\psi}{K}+\frac{4\eta}{K}+\frac{3\eta\sigma^2}{K}+\frac{6\eta}{K^2}\right)\\ &+\frac{4}{\eta}\left(\frac{\Delta_\psi}{K}+\frac{4\eta}{K}+\frac{3\eta\sigma^2}{K}+\frac{6\eta}{K^2}\right)+\frac{20+24\sigma^2}{K}+\frac{48}{K^2}\\ \leq& \frac{60\Delta_\psi L+80+69\sigma^2}{K}+\frac{138}{K^2}\leq \mathcal{O}\left(\frac{1}{K}\right).\end{aligned} \tag{113}$$

We choose $K=C\epsilon^{-1}$, where $C>0$ is a constant. Then, based on the above inequality, for C large enough, our PROX-ZO-SPIDER-Coord achieves $\mathbb{E}\|G(\mathbf{x}^\zeta,\nabla f(\mathbf{x}^\zeta),\eta)\|^2\leq\epsilon$, and the total number of function queries is

$$\begin{aligned}\left\lceil\frac{K}{q}\right\rceil Kd+K|\mathcal{S}_2|d &\leq Kd+K^{3/2}d+K^{3/2}d+Kd=2K^{3/2}d+2Kd\leq\mathcal{O}(K^{3/2}d)\\ &\leq\mathcal{O}(d\epsilon^{-3/2})\leq\mathcal{O}(\epsilon^{-1}n^{1/2}d),\end{aligned} \tag{114}$$

where the last two inequalities follow from the assumption that $n>k\geq C\epsilon^{-1}$.

Combining (111) and (114) in these two cases finishes the proof.

I. Proof for ZO-SVRG-Coord-Rand-C

Based on (3) in Lemma 5, we first establish the following key lemma.

Lemma 9. *Under Assumption 1, we have, for any $qk_0\leq k\leq\min\{q(k_0+1)-1,qh\}$, $k_0=0,...,h$,*

$$\begin{aligned}\mathbb{E}\|\mathbf{v}^k\|^2\leq&18dL\mathbb{E}(f(\mathbf{x}^k)-f(\mathbf{x}^*_\beta)+f(\mathbf{x}^{qk_0})-f(\mathbf{x}^*_\beta))\\ &+9L^2d\delta^2+\frac{45\beta^2L^2d^2}{4}+\frac{27I(|\mathcal{S}|<n)}{|\mathcal{S}|}\left(2L^2d\delta^2+\sigma^2\right)\end{aligned}$$

where $\mathbf{x}^*_\beta=\arg\min_{\mathbf{x}}f_\beta(\mathbf{x})$.

Proof. To simplify notation, we define

$$\widehat{\nabla}f_{i_k}(\mathbf{x})=\frac{d(f_{i_k}(\mathbf{x}+\beta\mathbf{u}^k)-f_{i_k}(\mathbf{x}))}{\beta}\mathbf{u}^k.$$

Based on the definition of $\mathbf{v}^k$ in ZO-SVRG-Coord-Rand-C, we have

$$\begin{aligned}\mathbb{E}\|\mathbf{v}^k\|^2=&\mathbb{E}\|\widehat{\nabla}f_{i_k}(\mathbf{x}^k)-\widehat{\nabla}f_{i_k}(\mathbf{x}^{qk_0})+\mathbf{v}^{qk_0}\|^2\\ \overset{\text{(i)}}{\leq}&3\mathbb{E}\|\widehat{\nabla}f_{i_k}(\mathbf{x}^k)-\widehat{\nabla}f_{i_k}(\mathbf{x}^*_\beta)\|^2+3\mathbb{E}\|\widehat{\nabla}f_{i_k}(\mathbf{x}^{qk_0})-\widehat{\nabla}f_{i_k}(\mathbf{x}^*_\beta)-(\nabla f_\beta(\mathbf{x}^{qk_0})-\nabla f_\beta(\mathbf{x}^*_\beta))\|^2\\ &+3\mathbb{E}\|\mathbf{v}^{qk_0}-\nabla f_\beta(\mathbf{x}^{qk_0})\|^2\\ \overset{\text{(ii)}}{\leq}&3\mathbb{E}\|\widehat{\nabla}f_{i_k}(\mathbf{x}^k)-\widehat{\nabla}f_{i_k}(\mathbf{x}^*_\beta)\|^2+3\mathbb{E}\|\widehat{\nabla}f_{i_k}(\mathbf{x}^{qk_0})-\widehat{\nabla}f_{i_k}(\mathbf{x}^*_\beta)\|^2+3\mathbb{E}\|\mathbf{v}^{qk_0}-\nabla f_\beta(\mathbf{x}^{qk_0})\|^2\\ \overset{\text{(iii)}}{\leq}&9d\mathbb{E}\|\nabla f_{i_k}(\mathbf{x}^k)-\nabla f_{i_k}(\mathbf{x}^*_\beta)\|^2+\frac{9L^2d^2\beta^2}{2}+9d\mathbb{E}\|\nabla f_{i_k}(\mathbf{x}^{qk_0})-\nabla f_{i_k}(\mathbf{x}^*_\beta)\|^2+\frac{9L^2d^2\beta^2}{2}\\ &+3\mathbb{E}\|\mathbf{v}^{qk_0}-\nabla f_\beta(\mathbf{x}^{qk_0})\|^2\\ \overset{\text{(iv)}}{\leq}&18dL\mathbb{E}(f_{i_k}(\mathbf{x}^k)-f_{i_k}(\mathbf{x}^*_\beta))+18dL\mathbb{E}(f_{i_k}(\mathbf{x}^{qk_0})-f_{i_k}(\mathbf{x}^*_\beta))+9L^2d^2\beta^2+3\mathbb{E}\|\mathbf{v}^{qk_0}-\nabla f_\beta(\mathbf{x}^{qk_0})\|^2\\ =&18dL\mathbb{E}(f(\mathbf{x}^k)-f(\mathbf{x}^*_\beta)+f(\mathbf{x}^{qk_0})-f(\mathbf{x}^*_\beta))+9L^2d^2\beta^2+3\mathbb{E}\|\mathbf{v}^{qk_0}-\nabla f_\beta(\mathbf{x}^{qk_0})\|^2\\ \overset{\text{(v)}}{\leq}&18dL\mathbb{E}(f(\mathbf{x}^k)-f(\mathbf{x}^*_\beta)+f(\mathbf{x}^{qk_0})-f(\mathbf{x}^*_\beta))+9L^2d\delta^2+\frac{45\beta^2L^2d^2}{4}\\ &+\frac{27I(|\mathcal{S}|<n)}{|\mathcal{S}|}\left(2L^2d\delta^2+\sigma^2\right)\end{aligned} \tag{115}$$

where (i) follows from the equality $\|\mathbf{a}+\mathbf{b}+\mathbf{c}\|^2 \leq 3(\|\mathbf{a}\|^2+\|\mathbf{b}\|^2+\|\mathbf{c}\|^2)$ and the fact that $\nabla f_\beta(\mathbf{x}_\beta^*)=0$, (ii) follows from the inequality $\mathbb{E}\|\mathbf{a}-\mathbb{E}(\mathbf{a})\|^2 \leq \mathbb{E}\|\mathbf{a}\|^2$, (iii) follows from item (3) in Lemma 5, (iv) follows from Lemma 5 in (Reddi et al., 2016a) that for convex and smooth function $f_{i_k}(\cdot)$,

$$\|\nabla f_{i_k}(\mathbf{x})-\nabla f_{i_k}(\mathbf{y})\|^2 \leq 2L(f_{i_k}(\mathbf{x})-f_{i_k}(\mathbf{y})-\langle \nabla f_{i_k}(\mathbf{x}), \mathbf{x}-\mathbf{y}\rangle),$$

and (v) follows from (30). □

Based on Lemma 5, we provide the following useful lemma as follows.

Lemma 10. *let Assumption 1 hold, and define the quantity*

$$\begin{aligned}\lambda =&\eta\left(9L^2d\delta^2+\frac{45\beta^2L^2d^2}{4}+\frac{27I(|\mathcal{S}|<n)}{|\mathcal{S}|}\left(2L^2d\delta^2+\sigma^2\right)\right)\\&+2\beta^2L+2\sqrt{3L^2d\delta^2+\frac{3\beta^2L^2d^2}{4}}\Gamma,\end{aligned} \tag{116}$$

where $\Gamma=\max_{0\leq k\leq K}\mathbb{E}\|\mathbf{x}^k-\mathbf{x}_\beta^*\|$. *Then, ZO-SVRG-Coord-Rand-C satisfies*

$$\mathbb{E}\big(f(\mathbf{x}^K)-f(\mathbf{x}^*)\big) \leq \alpha^h(f(\mathbf{x}^0)-f(\mathbf{x}_\beta^*)-\Delta)+\Delta+\beta^2L, \tag{117}$$

where $\mathbf{x}^*=\arg\min_{\mathbf{x}} f(\mathbf{x})$, $\alpha=18d\eta^2L/(2\eta-18d\eta^2L)$ *and* Δ *is given by*

$$\Delta=\frac{\Gamma^2/q+\eta\lambda}{2\eta-18d\eta^2L}\left(1+\frac{18d\eta^2L}{2\eta-36d\eta^2L}\right).$$

Proof. For $qm\leq k\leq q(m+1)-1, m=0,....,h-1$, we obtain the following sequence of inequalities

$$\begin{aligned}\mathbb{E}\|\mathbf{x}^{k+1}-\mathbf{x}_\beta^*\| =&\eta^2\mathbb{E}\|\mathbf{v}^k\|^2+\mathbb{E}\|\mathbf{x}^k-\mathbf{x}_\beta^*\|^2-2\eta\mathbb{E}\langle\mathbf{v}^k,\mathbf{x}^k-\mathbf{x}_\beta^*\rangle\\
=&\eta^2\mathbb{E}\|\mathbf{v}^k\|^2+\mathbb{E}\|\mathbf{x}^k-\mathbf{x}_\beta^*\|^2-2\eta\mathbb{E}\langle\nabla f_\beta(\mathbf{x}^k)-\nabla f_\beta(\mathbf{x}^{qm})+\mathbf{v}^{qm},\mathbf{x}^k-\mathbf{x}_\beta^*\rangle\\
=&\eta^2\mathbb{E}\|\mathbf{v}^k\|^2+\mathbb{E}\|\mathbf{x}^k-\mathbf{x}_\beta^*\|^2-2\eta\mathbb{E}\langle\nabla f_\beta(\mathbf{x}^k),\mathbf{x}^k-\mathbf{x}_\beta^*\rangle+2\eta\mathbb{E}\langle\nabla f_\beta(\mathbf{x}^{qm})-\mathbf{v}^{qm},\mathbf{x}^k-\mathbf{x}_\beta^*\rangle\\
\overset{\text{(i)}}{\leq}&\eta^2\mathbb{E}\|\mathbf{v}^k\|^2+\mathbb{E}\|\mathbf{x}^k-\mathbf{x}_\beta^*\|^2-2\eta\mathbb{E}(f_\beta(\mathbf{x}^k)-f_\beta(\mathbf{x}_\beta^*))+2\eta\mathbb{E}\|\nabla f_\beta(\mathbf{x}^{qm})-\mathbf{v}^{qm}\|\|\mathbf{x}^k-\mathbf{x}_\beta^*\|\\
\overset{\text{(ii)}}{\leq}&\mathbb{E}\|\mathbf{x}^k-\mathbf{x}_\beta^*\|^2+18d\eta^2L\mathbb{E}(f(\mathbf{x}^k)-f(\mathbf{x}_\beta^*)+f(\mathbf{x}^{qm})-f(\mathbf{x}_\beta^*))-2\eta\mathbb{E}(f_\beta(\mathbf{x}^k)-f_\beta(\mathbf{x}_\beta^*))\\
&+2\eta\sqrt{3L^2d\delta^2+\frac{3\beta^2L^2d^2}{4}}\Gamma+\eta^2\left(9L^2d\delta^2+\frac{45\beta^2L^2d^2}{4}+\frac{27I(|\mathcal{S}|<n)}{|\mathcal{S}|}\left(2L^2d\delta^2+\sigma^2\right)\right)\\
\overset{\text{(iii)}}{\leq}&\mathbb{E}\|\mathbf{x}^k-\mathbf{x}_\beta^*\|^2+18d\eta^2L\mathbb{E}(f(\mathbf{x}^k)-f(\mathbf{x}_\beta^*)+f(\mathbf{x}^{qm})-f(\mathbf{x}_\beta^*))-2\eta\mathbb{E}(f(\mathbf{x}^k)-f(\mathbf{x}_\beta^*))\\
&+2\eta\beta^2L+2\eta\sqrt{3L^2d\delta^2+\frac{3\beta^2L^2d^2}{4}}\Gamma\\
&+\eta^2\left(9L^2d\delta^2+\frac{45\beta^2L^2d^2}{4}+\frac{27I(|\mathcal{S}|<n)}{|\mathcal{S}|}\left(2L^2d\delta^2+\sigma^2\right)\right)\end{aligned} \tag{118}$$

where (i) follows from the convexity of $f_\beta(\cdot)$ (see (c) of Lemma 4.1 in Gao et al. 2014), (ii) follows from Lemma 9, (iii) follows from item (1) in Lemma 5. Then, telescoping (118) over k from qm to $q(m+1)-1$, we obtain

$$\begin{aligned}(2\eta-18d\eta^2L)\sum_{k=qm}^{q(m+1)-1}\mathbb{E}\left(f(\mathbf{x}^k)-f(\mathbf{x}_\beta^*)\right) \leq&\mathbb{E}(\|\mathbf{x}^{qm}-\mathbf{x}_\beta^*\|^2-\|\mathbf{x}^{q(m+1)}-\mathbf{x}_\beta^*\|^2)+q\eta\lambda\\
&+18dq\eta^2L\mathbb{E}(f(\mathbf{x}^{qm})-f(\mathbf{x}_\beta^*)).\end{aligned} \tag{119}$$

Based on ZO-SVRG-Coord-Rand-C, we have

$$\mathbb{E}f(\mathbf{x}^{q(m+1)})=\frac{1}{q}\sum_{k=qm}^{q(m+1)-1}\mathbb{E}f(\mathbf{x}^k),$$

which, in conjunction with (119), implies that

$$(2\eta - 18d\eta^2 L)\mathbb{E}\big(f(\mathbf{x}^{q(m+1)}) - f(\mathbf{x}_\beta^*)\big) \leq \frac{\Gamma^2}{q} + \eta\lambda + 18d\eta^2 L\mathbb{E}(f(\mathbf{x}^{qm}) - f(\mathbf{x}_\beta^*)). \tag{120}$$

Then, based on the selection of α and Δ in Theorem 10, we obtain from (120) that

$$\mathbb{E}\big(f(\mathbf{x}^{q(m+1)}) - f(\mathbf{x}_\beta^*)\big) - \Delta \leq \alpha\left(\mathbb{E}\big(f(\mathbf{x}^{qm}) - f(\mathbf{x}_\beta^*)\big) - \Delta\right).$$

Telescoping the above inequality over m from 0 to $h-1$, we obtain

$$\mathbb{E}\big(f(\mathbf{x}^K) - f(\mathbf{x}_\beta^*)\big) - \Delta \leq \alpha^h(\mathbb{E}\big(f(\mathbf{x}^0) - f(\mathbf{x}_\beta^*)\big) - \Delta). \tag{121}$$

Based on (1) in Lemma 5 and the definition of $\mathbf{x}_\beta^*$, we have

$$f(\mathbf{x}^*) - f(\mathbf{x}_\beta^*) \geq f_\beta(\mathbf{x}^*) - \frac{\beta^2 L}{2} - f_\beta(\mathbf{x}_\beta^*) - \frac{\beta^2 L}{2} \geq -\beta^2 L,$$

which, in conjunction with (121), yields

$$\mathbb{E}\big(f(\mathbf{x}^K) - f(\mathbf{x}^*)\big) - \Delta - \beta^2 L \leq \alpha^h(\mathbb{E}\big(f(\mathbf{x}^0) - f(\mathbf{x}_\beta^*)\big) - \Delta). \tag{122}$$

Then, the proof is complete. □

I.1. Proof of Theorem 6

Based on Lemmas 5 and 10, we now prove Theorem 6. We prove two cases with $n \leq \lceil c_s/\epsilon \rceil$ and $n < \lceil c_s/\epsilon \rceil$, separately.

First suppose that $n \leq \lceil c_s/\epsilon \rceil$, and thus $|\mathcal{S}| = n$. Then, applying the parameters selected in Corollary 6 in Theorem 10, we obtain $\alpha = 1/2$ and

$$\lambda = \frac{\epsilon^2}{3c_\delta^2 dL} + \frac{5\epsilon^2}{12c_\beta^2 dL} + \frac{2\epsilon^2}{c_\beta^2 d^2 L} + 2\Gamma\sqrt{\frac{3\epsilon^2}{c_\delta^2} + \frac{3\epsilon^2}{4c_\beta^2}},$$
$$\Delta = \frac{3}{2}\left(\frac{\Gamma^2}{q\eta} + \lambda\right) \leq \frac{3}{2}\left(\frac{27L\Gamma^2\epsilon}{c_q} + \frac{\epsilon^2}{3c_\delta^2 dL} + \frac{5\epsilon^2}{12c_\beta^2 dL} + \frac{2\epsilon^2}{c_\beta^2 d^2 L} + 2\Gamma\sqrt{\frac{3\epsilon^2}{c_\delta^2} + \frac{3\epsilon^2}{4c_\beta^2}}\right)$$

which, in conjunction with (117), implies that

$$\begin{aligned}\mathbb{E}\left(f(\mathbf{x}^K) - f(\mathbf{x}^*)\right) \leq & \frac{(f(\mathbf{x}^0) - f(\mathbf{x}_\beta^*))\epsilon}{c_h} + \frac{3}{2}\left(\frac{27L\Gamma^2\epsilon}{c_q} + \frac{\epsilon^2}{3c_\delta^2 dL} + \frac{5\epsilon^2}{12c_\beta^2 dL} + \frac{2\epsilon^2}{c_\beta^2 d^2 L} + 2\Gamma\sqrt{\frac{3\epsilon^2}{c_\delta^2} + \frac{3\epsilon^2}{4c_\beta^2}}\right) \\ & + \frac{\epsilon^2}{c_\beta^2 d^2 L}.\end{aligned} \tag{123}$$

For $c_h, c_q, c_\beta, c_\delta$ large enough, we obtain from (123) that $\mathbb{E}\left(f(\mathbf{x}^K) - f(\mathbf{x}^*)\right) \leq \epsilon$, and the number of function queries required by ZO-SVRG-Coord-Rand-C is at most

$$\begin{aligned}\left\lceil \frac{K}{q} \right\rceil nd + K = hnd + hq &= \log_2(c_h/\epsilon)nd + \log_2(c_h/\epsilon)c_q d/\epsilon \leq \mathcal{O}\left(d(n + 1/\epsilon)\log(1/\epsilon)\right) \\ &\leq \mathcal{O}(d\min\{n, 1/\epsilon\}\log(1/\epsilon)),\end{aligned}$$

where the last inequality follows from the assumption that $n \leq \lceil c_s/\epsilon \rceil$.

Next, suppose $n > \lceil c_s/\epsilon \rceil$, and thus $|\mathcal{S}| = \lceil c_s/\epsilon \rceil$. Then, we obtain

$$\lambda = \frac{\epsilon^2}{3c_\delta^2 dL} + \frac{5\epsilon^2}{12c_\beta^2 dL} + \frac{2\epsilon^2}{c_\beta^2 d^2 L} + 2\Gamma\sqrt{\frac{3\epsilon^2}{c_\delta^2} + \frac{3\epsilon^2}{4c_\beta^2}} + \frac{2\epsilon^3 + \sigma^2\epsilon}{c_s dL}$$
$$\Delta = \frac{3}{2}\left(\frac{\Gamma^2}{q\eta} + \lambda\right) \leq \frac{3}{2}\left(\frac{27L\Gamma^2\epsilon}{c_q} + \frac{\epsilon^2}{3c_\delta^2 dL} + \frac{5\epsilon^2}{12c_\beta^2 dL} + \frac{2\epsilon^2}{c_\beta^2 d^2 L} + 2\Gamma\sqrt{\frac{3\epsilon^2}{c_\delta^2} + \frac{3\epsilon^2}{4c_\beta^2}} + \frac{2\epsilon^3 + \sigma^2\epsilon}{c_s dL}\right)$$

which, in conjunction with (117), implies that

$$\mathbb{E}\left(f(\mathbf{x}^K)-f(\mathbf{x}^*)\right) \leq \frac{3}{2}\left(\frac{27L\Gamma^2\epsilon}{c_q}+\frac{\epsilon^2}{3c_\delta^2 dL}+\frac{5\epsilon^2}{12c_\beta^2 dL}+\frac{2\epsilon^2}{c_\beta^2 d^2 L}+2\Gamma\sqrt{\frac{3\epsilon^2}{c_\delta^2}+\frac{3\epsilon^2}{4c_\beta^2}}+\frac{2\epsilon^3+\sigma^2\epsilon}{c_s dL}\right) + \frac{(f(\mathbf{x}^0)-f(\mathbf{x}_\beta^*))\epsilon}{c_h}+\frac{\epsilon^2}{c_\beta^2 d^2 L}. \tag{124}$$

For $c_h, c_q, c_\beta, c_\delta, c_s$ large enough, we obtain from (123) that $\mathbb{E}\left(f(\mathbf{x}^K)-f(\mathbf{x}^*)\right)\leq\epsilon$, and the number of function queries required by our ZO-SVRG-Coord-Rand-C is

$$\left\lceil\frac{K}{q}\right\rceil|\mathcal{S}|d+K=h|\mathcal{S}|d+hq\leq\mathcal{O}\left(d(1/\epsilon)\log(1/\epsilon)\right)\leq\mathcal{O}(d\min\{n,1/\epsilon\}\log(1/\epsilon)),$$

where the last inequality follows from the assumption that $n>\lceil c_s/\epsilon\rceil$.

J. Proofs for ZO-SPIDER-Coord-C

J.1. Auxiliary Lemma

To prove the main theorem, we first establish two useful lemmas.

Lemma 11. *For any $qk_0\leq m\leq q(k_0+1), k_0=0,....,h-1$, we have*

$$\mathbb{E}\|\mathbf{v}^k-\hat{\nabla}_{\text{coord}}f(\mathbf{x}^k)\|^2 \leq 6(k-qk_0)L^2d\delta^2+3\sum_{m=qk_0+1}^{k}\mathbb{E}\|\nabla f_{i_m}(\mathbf{x}^m)-\nabla f_{i_m}(\mathbf{x}^{m-1})\|^2 + \frac{3I(|\mathcal{S}|<n)}{|\mathcal{S}|}\left(2L^2d\delta^2+\sigma^2\right).$$

where we define $\sum_{t=qk_0}^{qk_0-1}\mathbb{E}\|\mathbf{v}^t\|^2=0$ for simplicity.

Proof. Using an approach similar to (76) in Lemma 6 with $|\mathcal{S}_2|=1$, we obtain, for $qk_0+1\leq m\leq k$,

$$\mathbb{E}\|\mathbf{v}^m-\hat{\nabla}_{\text{coord}}f(\mathbf{x}^m)\|^2\leq 6L^2d\delta^2+3\|\nabla f_{i_m}(\mathbf{x}^m)-\nabla f_{i_m}(\mathbf{x}^{m-1})\|^2+\mathbb{E}\|\mathbf{v}^{m-1}-\hat{\nabla}_{\text{coord}}f(\mathbf{x}^{m-1})\|^2.$$

Telescoping the above inequality over m from qk_0+1 to k yields

$$\mathbb{E}\|\mathbf{v}^k-\hat{\nabla}_{\text{coord}}f(\mathbf{x}^k)\|^2 \leq 6(k-qk_0)L^2d\delta^2+3\sum_{m=qk_0+1}^{k}\|\nabla f_{i_m}(\mathbf{x}^m)-\nabla f_{i_m}(\mathbf{x}^{m-1})\|^2 + \mathbb{E}\|\mathbf{v}^{qk_0}-\hat{\nabla}_{\text{coord}}f(\mathbf{x}^{qk_0})\|^2$$

which, in conjunction with Lemma 4, finishes the proof. □

Lemma 12. *For any $qk_0\leq m\leq q(k_0+1), k_0=0,....,h-1$, we have*

$$\sum_{m=qk_0+1}^{k}\mathbb{E}\|\hat{\nabla}_{\text{coord}}f_{i_m}(\mathbf{x}^m)-\hat{\nabla}_{\text{coord}}f_{i_m}(\mathbf{x}^{m-1})\|^2\leq\frac{L\eta}{2-L\eta}\mathbb{E}\|\mathbf{v}^{qk_0}\|^2. \tag{125}$$

Proof. Define a smoothing function of $f(\mathbf{x})$ with regard to its i^{th} coordinate as $f_{i,\delta}(\mathbf{x})=\mathbb{E}_{\mathbf{v}\sim\text{U}(-\delta,\delta)}(f(\mathbf{x}+\mathbf{v}\mathbf{e}_i))$, where $\text{U}(-\delta,\delta)$ denotes the uniform distribution over the range $[-\delta,\delta]$. Then, based on Lemma 6 in (Lian et al., 2016), the function $f_{i,\delta}(\mathbf{x})$ has the following three useful properties:

(1) $\mathbf{e}_i\mathbf{e}_i^T\nabla f_{i,\delta}(\mathbf{x})=\frac{1}{2\delta}\left(f(\mathbf{x}+\delta\mathbf{e}_i)-f(\mathbf{x}+\delta\mathbf{e}_i)\right)\mathbf{e}_i$

(2) If $f(\mathbf{x})$ has the L-Lipschitz gradient, then $f_{i,\delta}(\mathbf{x})$ also has the L-Lipschitz gradient.

(3) If $f(\mathbf{x})$ is convex, then $f_{i,\delta}(\mathbf{x})$ is convex.

Based the above preliminaries, we next prove Lemma 12. Recall from ZO-SPIDER-Coord-C that

$$\mathbf{v}^m = \hat{\nabla}_{\text{coord}} f_{i_m}(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f_{i_m}(\mathbf{x}^{m-1}) + \mathbf{v}^{k-1}, \tag{126}$$

where we recall that for $\mathbf{x} = \mathbf{x}^m$ and $\mathbf{x}^{m-1}$

$$\hat{\nabla}_{\text{coord}} f_{i_m}(\mathbf{x}^m) = \sum_{i=1}^{d} \frac{1}{2\delta}(f_{i_m}(\mathbf{x} + \delta \mathbf{e}_i) - f_{i_m}(\mathbf{x} - \delta \mathbf{e}_i))\mathbf{e}_i = \sum_{i=1}^{d} \mathbf{e}_i \mathbf{e}_i^T \nabla f_{i_m,i,\delta}(\mathbf{x}). \tag{127}$$

Then, based on (126), we have

$$\begin{aligned} \|\mathbf{v}^m\|^2 =& \|\mathbf{v}^{m-1}\|^2 + \|\hat{\nabla}_{\text{coord}} f_{i_m}(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f_{i_m}(\mathbf{x}^{m-1})\|^2 \\ &+ 2\underbrace{\langle \hat{\nabla}_{\text{coord}} f_{i_m}(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f_{i_m}(\mathbf{x}^{m-1}), \mathbf{v}^{m-1} \rangle}_{\text{(I)}}. \end{aligned} \tag{128}$$

We next upper-bound the term (I) in the above equation using the convexity of function $f_{i_k}(\cdot)$. In specific, we have

$$\begin{aligned} \text{(I)} =& -\frac{1}{\eta} \left\langle \sum_{i=1}^{d} \mathbf{e}_i \mathbf{e}_i^T \left(\nabla f_{i_m,i,\delta}(\mathbf{x}^m) - \nabla f_{i_m,i,\delta}(\mathbf{x}^{m-1}) \right), \mathbf{x}^m - \mathbf{x}^{m-1} \right\rangle \\ \overset{\text{(i)}}{=}& -\sum_{i=1}^{d} \frac{1}{\eta} \left\langle \nabla f_{i_m,i,\delta}(\mathbf{e}_i \mathbf{e}_i^T \mathbf{x}^m) - \nabla f_{i_m,i,\delta}(\mathbf{e}_i \mathbf{e}_i^T \mathbf{x}^{m-1}), \mathbf{e}_i \mathbf{e}_i^T \mathbf{x}^m - \mathbf{e}_i \mathbf{e}_i^T \mathbf{x}^{m-1} \right\rangle \\ \overset{\text{(ii)}}{\leq}& -\sum_{i=1}^{d} \frac{1}{L\eta} \left\| \nabla f_{i_m,i,\delta}(\mathbf{e}_i \mathbf{e}_i^T \mathbf{x}^m) - \nabla f_{i_m,i,\delta}(\mathbf{e}_i \mathbf{e}_i^T \mathbf{x}^{m-1}) \right\|^2 \\ =& -\sum_{i=1}^{d} \frac{1}{L\eta} \left\| \mathbf{e}_i \mathbf{e}_i^T \left(\nabla f_{i_m,i,\delta}(\mathbf{x}^m) - \nabla f_{i_m,i,\delta}(\mathbf{x}^{m-1}) \right) \right\|^2 \\ =& -\frac{1}{L\eta} \left\| \hat{\nabla}_{\text{coord}} f_{i_m}(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f_{i_m}(\mathbf{x}^{m-1}) \right\|^2. \end{aligned} \tag{129}$$

where (i) follows from the definition of $\mathbf{e}_i$, (ii) follows from the convexity of $f_{i_m,i,\delta}(\cdot)$ and Theorem 2.1.5 in (Nesterov, 2013), and the last inequality follows from the definition of $\mathbf{e}_i$ and the ℓ_2-norm. Combining (128) and (129) implies that

$$\|\mathbf{v}^m\|^2 = \|\mathbf{v}^{m-1}\|^2 + \left(1 - \frac{2}{L\eta}\right) \|\hat{\nabla}_{\text{coord}} f_{i_m}(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f_{i_m}(\mathbf{x}^{m-1})\|^2$$

Telescoping the above inequality over m from $qk_0 + 1$ to k and taking the expectation, we finish the proof. □

Based on Lemmas 11 and 12, we next prove the following useful lemma.

Lemma 13. *Under Assmption 1, we define*

$$\lambda = 3qL^2 d\delta^2 + \frac{6L\eta}{2 - L\eta} L^2 d\delta^2 + \frac{3I(|\mathcal{S}| < n)}{|\mathcal{S}|} \left(2L^2 d\delta^2 + \sigma^2\right). \tag{130}$$

Then, our ZO-SPIDER-Coord-C satisfies

$$\mathbb{E}\|\nabla f(\mathbf{x}^K)\|^2 \leq \alpha^h \mathbb{E}\|\nabla f(\mathbf{x}^0)\|^2 + \frac{1 - \alpha^h}{1 - \alpha} \Delta,$$

with the parameters satisfying $\alpha = 6L(\eta + 2L\eta^2)(2 - L\eta)^{-1}(1/2 - L\eta)^{-1}$ *and*

$$\Delta = \frac{\Gamma}{q(\frac{\eta}{2} - L\eta^2)} + \frac{1 + 2L\eta}{\frac{1}{2} - L\eta} \lambda, \tag{131}$$

where $\Gamma = \max_{0 \leq k \leq h} \{\mathbb{E}\left(f(\mathbf{x}^{qk}) - f(\mathbf{x}^*)\right)\}$ *with* $\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$.

Proof. Since $f(\cdot)$ has the L-Lipschitz gradient, we have, for $qk_0 \leq m \leq q(k_0+1), k_0 = 0, ..., h-1$,

$$\begin{aligned}
f(\mathbf{x}^{m+1}) &\leq f(\mathbf{x}^m) + \langle \nabla f(\mathbf{x}^m), \mathbf{x}^{m+1} - \mathbf{x}^m \rangle + \frac{L}{2}\|\mathbf{x}^{m+1} - \mathbf{x}^m\|^2 \\
&= f(\mathbf{x}^m) - \eta \langle \mathbf{v}^m - \nabla f(\mathbf{x}^m), \nabla f(\mathbf{x}^m) \rangle - \eta \|\nabla f(\mathbf{x}^m)\|^2 + \frac{L\eta^2}{2}\|\mathbf{v}^m\|^2 \\
&\leq f(\mathbf{x}^m) + \frac{\eta}{2}\|\nabla f(\mathbf{x}^m) - \mathbf{v}^m\|^2 - \frac{\eta}{2}\|\nabla f(\mathbf{x}^m)\|^2 + L\eta^2\|\mathbf{v}^m - \nabla f(\mathbf{x}^m)\|^2 + L\eta^2\|\nabla f(\mathbf{x}^m)\|^2 \\
&= f(\mathbf{x}^m) + \left(\frac{\eta}{2} + L\eta^2\right)\|\nabla f(\mathbf{x}^m) - \mathbf{v}^m\|^2 - \left(\frac{\eta}{2} - L\eta^2\right)\|\nabla f(\mathbf{x}^m)\|^2, \\
&\leq f(\mathbf{x}^m) + \left(\eta + 2L\eta^2\right)\left(\|\hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) - \mathbf{v}^m\|^2 + \|\nabla f(\mathbf{x}^m) - \hat{\nabla}_{\text{coord}} f(\mathbf{x}^m)\|^2\right) - \left(\frac{\eta}{2} - L\eta^2\right)\|\nabla f(\mathbf{x}^m)\|^2.
\end{aligned}$$

Taking expectation over the above inequality and using Lemmas 3, 11 and 12, we have

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}^{m+1}) \leq& \mathbb{E}f(\mathbf{x}^m) + \left(\eta + 2L\eta^2\right)\left(\mathbb{E}\|\hat{\nabla}_{\text{coord}} f(\mathbf{x}^m) - \mathbf{v}^m\|^2 + L^2 d\delta^2\right) - \left(\frac{\eta}{2} - L\eta^2\right)\mathbb{E}\|\nabla f(\mathbf{x}^m)\|^2 \\
\leq& \mathbb{E}f(\mathbf{x}^m) + \left(\eta + 2L\eta^2\right)\Bigg(6(k - qk_0)L^2 d\delta^2 + L^2 d\delta^2 + 3\sum_{m=qk_0+1}^{k}\mathbb{E}\|\nabla f_{i_m}(\mathbf{x}^m) - \nabla f_{i_m}(\mathbf{x}^{m-1})\|^2 \\
&+ \frac{3I(|\mathcal{S}_1| < n)}{|\mathcal{S}_1|}\left(2L^2 d\delta^2 + \sigma^2\right)\Bigg) - \left(\frac{\eta}{2} - L\eta^2\right)\mathbb{E}\|\nabla f(\mathbf{x}^m)\|^2 \\
\overset{\text{(i)}}{\leq}& \mathbb{E}f(\mathbf{x}^m) + \left(\eta + 2L\eta^2\right)\left(6(k - qk_0)L^2 d\delta^2 + L^2 d\delta^2 + \frac{3L\eta}{2 - L\eta}\mathbb{E}\|\mathbf{v}^{qk_0}\|^2 + \frac{3I(|\mathcal{S}| < n)}{|\mathcal{S}|}\left(2L^2 d\delta^2 + \sigma^2\right)\right) \\
&- \left(\frac{\eta}{2} - L\eta^2\right)\mathbb{E}\|\nabla f(\mathbf{x}^m)\|^2
\end{aligned}$$

where (i) follows from Lemma 12. Noting that $\mathbf{v}^{qk_0} = \hat{\nabla}_{\text{coord}} f(\mathbf{x}^{qk_0})$ and telescoping the above inequality over m from qk_0 to $q(k_0+1) - 1$, we obtain

$$\begin{aligned}
&\sum_{m=qk_0}^{q(k_0+1)-1}\left(\frac{\eta}{2} - L\eta^2\right)\mathbb{E}\|\nabla f(\mathbf{x}^m)\|^2 \leq \mathbb{E}f(\mathbf{x}^{qk_0}) - \mathbb{E}f(\mathbf{x}^{q(k_0+1)}) \\
&+ \left(\eta + 2L\eta^2\right)\left(3q^2L^2 d\delta^2 + \frac{6qL\eta}{2 - L\eta}\mathbb{E}\|\nabla f(\mathbf{x}^{qk_0})\|^2 + \frac{6qL\eta}{2 - L\eta}L^2 d\delta^2 + \frac{3qI(|\mathcal{S}| < n)}{|\mathcal{S}|}\left(2L^2 d\delta^2 + \sigma^2\right)\right) \qquad (132)
\end{aligned}$$

Combining (130) with (132) implies that

$$\sum_{m=qk_0}^{q(k_0+1)-1}\left(\frac{\eta}{2} - L\eta^2\right)\mathbb{E}\|\nabla f(\mathbf{x}^m)\|^2 \leq \mathbb{E}f(\mathbf{x}^{qk_0}) - \mathbb{E}f(\mathbf{x}^*) + \left(\eta + 2L\eta^2\right)\left(\frac{6qL\eta}{2 - L\eta}\mathbb{E}\|\nabla f(\mathbf{x}^{qk_0})\|^2 + q\lambda\right),$$

which, in conjunction with the fact that $\mathbf{x}^{q(k_0+1)}$ is generated from $\{\mathbf{x}^{qk_0}, ..., \mathbf{x}^{q(k_0+1)-1}\}$ uniformly at random and (131), yields

$$\mathbb{E}\|\nabla f(\mathbf{x}^{q(k_0+1)})\|^2 \leq \alpha\mathbb{E}\|\nabla f(\mathbf{x}^{qk_0})\|^2 + \Delta.$$

Telescoping the above inequality over k_0 from 0 to $h-1$ yields

$$\mathbb{E}\|\nabla f(\mathbf{x}^K)\|^2 \leq \alpha^h\mathbb{E}\|\nabla f(\mathbf{x}^0)\|^2 + \frac{1 - \alpha^h}{1 - \alpha}\Delta, \qquad (133)$$

which finishes the proof. □

J.2. Proof of Theorem 7

Using Lemmas 11, 12 and 13, we prove Theorem 7. We prove two cases with $n \leq \lceil c_s/\epsilon \rceil$ and $n < \lceil c_s/\epsilon \rceil$, separately.

First suppose that $n \leq \lceil c_s/\epsilon \rceil$, and thus $|\mathcal{S}| = n$. Then, applying the parameters selected in Corollary 7 in Theorem 13, we obtain $\alpha \leq 1/2$ and $\Delta \leq \mathcal{O}(\epsilon/c_q)$ which, in conjunction with (117), implies that

$$\mathbb{E}\|\nabla f(\mathbf{x}^K)\|^2 \leq \mathcal{O}\left(\frac{\epsilon}{c_h} + \frac{\epsilon}{c_q}\right). \tag{134}$$

For c_h, c_q large enough, we obtain from (134) that $\mathbb{E}\|\nabla f(\mathbf{x}^K)\|^2 \leq \epsilon$, and the number of function queries required by our ZO-SPIDER-Coord-C is at most

$$\begin{aligned}\left\lceil \frac{K}{q} \right\rceil nd + Kd = hnd + hqd = \log_2(c_h/\epsilon)nd + \log_2(c_h/\epsilon)c_q d/\epsilon &\leq \mathcal{O}\left(d(n + 1/\epsilon)\log(1/\epsilon)\right) \\ &\leq \mathcal{O}(d\min\{n, 1/\epsilon\}\log(1/\epsilon)),\end{aligned}$$

where the last inequality follows from the assumption that $n \leq \lceil c_s/\epsilon \rceil$.

Next, suppose $n > \lceil c_s/\epsilon \rceil$, and thus $|\mathcal{S}| = \lceil c_s/\epsilon \rceil$. Then, we similarly obtain

$$\mathbb{E}\|\nabla f(\mathbf{x}^K)\|^2 \leq \mathcal{O}\left(\frac{\epsilon}{c_q} + \frac{\epsilon}{c_h} + \frac{\epsilon}{c_s}\right).$$

Then. for c_h, c_q, c_s large enough, we obtain from (123) that $\mathbb{E}\|\nabla f(\mathbf{x}^K)\|^2 \leq \epsilon$, and the number of function queries required by ZO-SPIDER-Coord-C is given by

$$\left\lceil \frac{K}{q} \right\rceil |\mathcal{S}|d + Kd = h|\mathcal{S}|d + hdq \leq \mathcal{O}\left(d(1/\epsilon)\log(1/\epsilon)\right) \leq \mathcal{O}(d\min\{n, 1/\epsilon\}\log(1/\epsilon)),$$

where the last inequality follows from the assumption that $n > \lceil c_s/\epsilon \rceil$.